

Revisiting Perceptron: Efficient and Label-Optimal Active Learning of Halfspaces

Songbai Yan^{*1} and Chicheng Zhang^{†1}

¹University of California, San Diego

February 21, 2017

Abstract

It has been a long-standing problem to efficiently learn a linear separator using as few labels as possible. In this work, we propose an efficient perceptron-based algorithm for actively learning homogeneous linear separators under uniform distribution. Under bounded noise, where each label is flipped with probability at most η , our algorithm achieves near-optimal $\tilde{O}\left(\frac{d}{(1-2\eta)^2} \log \frac{1}{\epsilon}\right)^1$ label complexity in time $\tilde{O}\left(\frac{d^2}{\epsilon(1-2\eta)^2}\right)$, and significantly improves over the best known result [Awasthi et al., 2016]. Under adversarial noise, where at most ν -fraction of labels can be flipped, our algorithm achieves near-optimal $\tilde{O}(d \log \frac{1}{\epsilon})$ label complexity in time $\tilde{O}\left(\frac{d^2}{\epsilon}\right)$, which is significantly better than the best known label complexity and time complexity in Awasthi et al. [2014].

1 Introduction

We study the problem of designing efficient noise-tolerant algorithms for actively learning homogeneous linear separators. We are given access to unlabeled samples and a labeling oracle that we can query for labels. The labels returned by the oracle may be noisy. We would like to find a computationally efficient algorithm to learn a linear separator that best classifies the data while making as few queries to the labeling oracle as possible.

Active learning arises naturally in many machine learning applications where unlabeled samples are abundant and cheap, but labeling requires human effort and is expensive. For those applications, one natural question is whether we can learn a good classifier while using as few labels as possible. Active learning addresses this question by allowing the learning algorithm to sequentially select examples to query for labels, and avoid requesting labels which are less informative, or can be inferred from previously-observed samples.

There has been a large body of work on the theory of active learning, showing sharp distribution-dependent label complexity bounds [Cohn et al., 1994, Freund et al., 1997, Dasgupta, 2005, Balcan et al., 2009, Hanneke, 2007a, Dasgupta et al., 2007, Koltchinskii, 2010, Beygelzimer et al., 2010, Wang, 2011, Hanneke, 2011, Zhang and Chaudhuri, 2014, Huang et al., 2015]. However, most of these general active learning algorithms need to solve empirical risk minimization problems, which are computationally hard in the presence of noise even for learning linear separators under uniform distribution [Klivans and Kothari, 2014].

^{*}yansongbai@eng.ucsd.edu

[†]chichengzhang@ucsd.edu

¹We use $\tilde{O}(f(\cdot)) = O(f(\cdot) \ln f(\cdot))$.

On the other hand, existing computationally efficient learning algorithms for linear separators are not label-efficient. A line of work considers efficient learning of linear separators with noise [Blum et al., 1998, Dunagan and Vempala, 2004, Kalai et al., 2008, Klivans et al., 2009, Awasthi et al., 2014, Daniely, 2015, Awasthi et al., 2015, 2016]. These algorithms have different degrees of noise tolerance (e.g. adversarial noise, malicious noise, random classification noise, bounded noise, etc), and run in time polynomial in $\frac{1}{\epsilon}$ and d . Some of them enjoy the feature of active learning Awasthi et al. [2014, 2015, 2016], but they do not achieve the sharpest label complexity bounds in contrast to those inefficient active learning algorithms.

Therefore, a natural open question is: is there any active learning halfspace algorithm that is computationally efficient, and has minimum label requirement? A variant of this problem has been posed as a COLT open problem by Monteleoni [2006]. In the realizable setting, Dasgupta et al. [2005], Balcan et al. [2007], Balcan and Long [2013] give efficient algorithms that have optimal label complexity of $\tilde{O}(d \ln \frac{1}{\epsilon})$ under some distributional assumptions. However, the question still remains open for the more challenging nonrealizable setting. Our paper gives an affirmative answer to this question under two noise settings: bounded noise condition and adversarial noise condition.

1.1 Our results

We propose a modified perceptron algorithm, ACTIVE-PERCEPTRON, for actively learning homogeneous linear separators under uniform distribution over unit sphere. It works under two noise settings: bounded noise and adversarial noise. In the η -bounded noise setting, the label of an example x is generated by $\text{sign}(u \cdot x)$ for some underlying linear separator u , and flipped with probability at most $\eta < \frac{1}{2}$. Our algorithm runs in time $\tilde{O}\left(\frac{d^2}{(1-2\eta)^3} \cdot \frac{1}{\epsilon} \ln \frac{1}{\epsilon}\right)$, and requires only $\tilde{O}\left(\frac{d}{(1-2\eta)^2} \cdot \ln \frac{1}{\epsilon}\right)$ labels. We show that this label complexity is *nearly optimal* by providing an almost matching information-theoretic lower bound of $\Omega\left(\frac{d}{(1-2\eta)^2} \cdot \ln \frac{1}{\epsilon}\right)$.

Our time complexity and label complexity significantly improve over the only known efficient algorithm proposed in Awasthi et al. [2016], whose time complexity and label complexity are both polynomials in d (the dimension) with a large unspecified degree. Our result also answers an open question by Dasgupta et al. [2005] on whether active perceptron algorithms can be modified to tolerate label noise.

Our main theorem on learning under bounded noise is shown below:

Theorem 1 (ACTIVE-PERCEPTRON under Bounded Noise). *Suppose Algorithm 1 has inputs labeling oracle \mathcal{O} that satisfies η -bounded noise condition with respect to u , initial halfspace v_0 , target error ϵ , confidence δ , then with probability $1 - \delta$: (1) The output halfspace v_k is such that $\mathbb{P}[\text{sign}(v_k \cdot x) \neq \text{sign}(u \cdot x)] \leq \epsilon$; (2) The total number of label queries to oracle \mathcal{O} is at most $\tilde{O}\left(\frac{d}{(1-2\eta)^2} \cdot \ln \frac{1}{\epsilon}\right)$; (3) The algorithm runs in time $\tilde{O}\left(\frac{d^2}{(1-2\eta)^3} \cdot \frac{1}{\epsilon} \ln \frac{1}{\epsilon}\right)$.*

Table 1 presents a comparison between our results and the results most closely related to our work.

In addition, we show that our algorithm also works for a more challenging setting, the ν -adversarial noise case, where a ν fraction of labels are flipped with respect to the labeling of underlying linear separator u . We show that our algorithm achieves an error of ϵ while tolerating a noise level of $\nu \leq O(\frac{\epsilon}{\ln \frac{1}{\delta} + \ln \ln \frac{1}{\epsilon}})$. Our algorithm runs in time $\tilde{O}(d^2 \cdot \frac{1}{\epsilon} \ln \frac{1}{\epsilon})$, and requires only $\tilde{O}(d \cdot \ln \frac{1}{\epsilon})$ labels. The best known result under this setting is Awasthi et al. [2014]. It works when $\nu \leq O(\epsilon)$, but requires $\tilde{O}(d^2)$ labels², and its time complexity bound is at least $\tilde{O}(d^3)$.

Our main theorem on adversarial noise is shown below:

Theorem 2 (ACTIVE-PERCEPTRON under Adversarial Noise). *Suppose Algorithm 1 has inputs labeling oracle \mathcal{O} that satisfies ν -adversarial noise condition with respect to u , initial halfspace v_0 , target error ϵ , confidence δ . Additionally $\nu < O(\frac{\epsilon}{\ln \frac{1}{\delta} + \ln \ln \frac{1}{\epsilon}})$. Then with probability $1 - \delta$: (1) The output halfspace v_k*

²A refined Rademacher complexity-based analysis shows that the label complexity can be sharpened to $\tilde{O}(d \ln \frac{1}{\epsilon})$, though this point is not made explicit in the paper.

Algorithm	Label Complexity	Time Complexity
Balcan et al. [2007]	$\tilde{O}(\frac{d}{(1-2\eta)^2} \ln \frac{1}{\epsilon})$	superpoly($d, \frac{1}{\epsilon}$) ³
Awasthi et al. [2016]	$\text{poly}(d, \frac{1}{1-2\eta}, \ln \frac{1}{\epsilon})$	$\text{poly}(d, \frac{1}{1-2\eta}, \frac{1}{\epsilon})$
Our Work	$\tilde{O}(\frac{d}{(1-2\eta)^2} \ln \frac{1}{\epsilon})$	$\tilde{O}(\frac{d^2}{(1-2\eta)^3} \cdot \frac{1}{\epsilon} \ln \frac{1}{\epsilon})$

Table 1: A comparison of algorithms for learning linear separators under uniform distribution, in η -bounded noise model.

Algorithm	Noise Tolerance ν	Label Complexity	Time Complexity
Zhang and Chaudhuri [2014]	$\Omega(\epsilon)$	$\tilde{O}(d \ln \frac{1}{\epsilon})$	superpoly($d, \frac{1}{\epsilon}$)
Awasthi et al. [2014]	$\Omega(\epsilon)$	$\tilde{O}(d^2 \ln \frac{1}{\epsilon})$	$\text{poly}(d, \frac{1}{\epsilon})$
Our Work	$\Omega(\frac{\epsilon}{\ln d + \ln \ln \frac{1}{\epsilon}})$	$\tilde{O}(d \ln \frac{1}{\epsilon})$	$\tilde{O}(d^2 \cdot \frac{1}{\epsilon} \ln \frac{1}{\epsilon})$

Table 2: A comparison of algorithms for learning linear separators under uniform distribution, in ν -adversarial noise model.

is such that $\mathbb{P}[\text{sign}(v_k \cdot x) \neq \text{sign}(u \cdot x)] \leq \epsilon$; (2) The total number of label queries to oracle \mathcal{O} is at most $\tilde{O}(d \cdot \ln \frac{1}{\epsilon})$. (3) The algorithm runs in time $\tilde{O}(d^2 \cdot \frac{1}{\epsilon} \ln \frac{1}{\epsilon})$.

Table 2 presents a comparison between our results and the results most closely related to our work.

1.2 Techniques

Our solution is based on a novel sampling criterion for perceptron-based active learning. For simplicity, suppose for the rest of this subsection that the data distribution is separable by a halfspace u . Consider the modified perceptron update rule [Motzkin and Schoenberg, 1954, Blum et al., 1998, Hampson and Kibler, 1999, Dasgupta et al., 2005, Crammer et al., 2010, Hanneke et al., 2015] over unit vector v_t :

$$v_{t+1} \leftarrow v_t - 2\mathbb{1}\{y_t \neq \text{sign}(v_t \cdot x_t)\} (v_t \cdot x_t) \cdot x_t \quad (1)$$

First, it can be seen that $\|v_{t+1}\| = \|v_t\| = 1$. Secondly, let θ_t denote the angle between v_t and u . It is shown in Dasgupta et al. [2005] that θ_t is monotonically nonincreasing, no matter how x_t is chosen. But how does the choice of x_t affect the convergence speed of the angle θ_t to 0?

Update rule (1) directly implies the following statement relating θ_{t+1} and θ_t :

$$\cos \theta_{t+1} = \cos \theta_t - 2\mathbb{1}\{y_t \neq \text{sign}(v_t \cdot x_t)\} (v_t \cdot x_t) \cdot (u \cdot x_t) \quad (2)$$

At time t , a natural sampling strategy could be sampling from $D|_{R_t}$, where $R_t = \{(x, y) : v_t \cdot x = b\}$, and $b \geq 0$ is a parameter to be specified. One should set b that maximizes the cost-efficiency of the update, that is, maximizing the expected increment of $\cos \theta_t$:

$$b_t = \arg \max_{b \in [0, 1]} \mathbb{E}_{(x, y) \sim D} \left[-2\mathbb{1}\{y \neq \text{sign}(v_t \cdot x)\} (u \cdot x) \mid v_t \cdot x = b \right]$$

It turns out that the setting of $b_t = \Theta(\frac{\theta_t}{\sqrt{d}})$ approximately maximizes the above quantity. Although we do not know θ_t exactly, we show that our epoch-based algorithm chooses near-optimal b_t 's in a frequent manner. This observation, in conjunction with martingale concentration bounds, together imply the convergence of the angle θ_t .

³The algorithm needs to minimize 0-1 loss, the best known method for which requires superpolynomial time.

2 Related work

Active Learning. Recent years have witnessed a wide range of success in both theory and applications of active learning; see surveys by Settles [2010], Hanneke et al. [2014], Dasgupta [2011]. On the theory side, many label-efficient active learning algorithms have been proposed and analyzed. An incomplete list includes Cohn et al. [1994], Freund et al. [1997], Dasgupta [2005], Balcan et al. [2009], Hanneke [2007a], Balcan et al. [2007], Dasgupta et al. [2007], Balcan et al. [2010], Beygelzimer et al. [2009], Hanneke [2009], Koltchinskii [2010], Hsu [2010], Beygelzimer et al. [2010], Wang [2011], Hanneke [2011], Ailon et al. [2014], Zhang and Chaudhuri [2014], Huang et al. [2015]. Most algorithms are *disagreement-based active learning* [Hanneke et al., 2014], and have suboptimal label complexity. Another crucial drawback is that, most of these algorithms need to solve empirical risk minimization problems, which is computationally hard in the presense of noise even for learning halfspaces under uniform distribution [Klivans and Kothari, 2014]. Recently Zhang and Chaudhuri [2014] proposes confidence-based active learning, where they can handle general data distribution, achieve statistical consistency, and have good label complexity bounds. Unfortunately the algorithms are still intractable.

Efficient Halfspace Learning. The problem of efficient learning of linear separators is one of the central problems in machine learning. In the realizable case, it is well known that standard linear programming will find a consistent hypothesis over data efficiently. In the general agnostic setting, the problem is much more challenging.

A series of papers have shown the hardness of learning separators with agnostic noise [Arora et al., 1993, Feldman et al., 2006, Guruswami and Raghavendra, 2009, Klivans and Kothari, 2014, Daniely, 2015]. The state-of-the-art result [Daniely, 2015] shows that under standard complexity-theoretic assumptions, there exists a data distribution, such that the best linear classifier has error $o(1)$, but no polynomial time algorithms can achieve an error at most $\frac{1}{2} - \frac{1}{d^\epsilon}$, even with improper learning. Klivans and Kothari [2014] shows that under standard assumptions, even if the unlabeled distribution of is Gaussian, any agnostic halfspace learning algorithm must run in time $d^{\Omega(\log \frac{1}{\epsilon})}$ to achieve an excess error of ϵ . These results indicate that, in order to have nontrivial guarantees on efficient learning linear separators with noise, one has to make additional assumptions over the data distribution.

On the other hand, positive results have been shown in various noise models under some restricted unlabeled data distributions (for example, uniform, isotropic log-concave). In the random classification noise model, Blum et al. [1998] gives the first efficient algorithm of learning linear separators that can tolerate such noise. However it is unclear how to apply the result to the more challenging bounded noise model (see below).

In the bounded noise model (also known as Massart noise model [Massart and Nédélec, 2006]), Awasthi et al. [2015] gives an efficient algorithm that learns linear separators can tolerate bounded noise of magnitude $\eta < 1.8 \times 10^{-6}$. Later, Awasthi et al. [2016] provides an efficient algorithm that combines the ideas of Awasthi et al. [2014] and Kalai et al. [2008], tolerating any $\eta < \frac{1}{2}$.

In the adversarial noise model, we assume that the optimal linear separator has error ν over data. Awasthi et al. [2014] shows an efficient algorithm that outputs a linear separator of error $O(\nu)$. Furthermore, Daniely [2015] gives a PTAS that outputs a classifier with error $(1 + \mu)\nu + \epsilon$, in time $O(\text{poly}(d^{\tilde{O}(\frac{1}{\mu^2})}, \frac{1}{\epsilon}))$.

Efficient Active Learning of Linear Separators. Despite considerable efforts, only a few label-efficient algorithms are computational-efficient even for learning linear separators under uniform distribution. In the realizable setting, Dasgupta et al. [2005], Balcan et al. [2007], Balcan and Long [2013] propose computation-efficient active learning algorithms which have an optimal label complexity of $\tilde{O}(d \log \frac{1}{\epsilon})$.

Since it is believed to be hard for learning linear separators in the general agnostic setting, it is natural to consider algorithms that work under more moderate noise conditions. A line of work considers specific noise models. For example, Dekel et al. [2012] gives an efficient algorithm for the setting that $\mathbb{P}[Y = 1|X = x] = \frac{1+u \cdot x}{2}$ where u is the optimal classifier. Agarwal [2013] studies generalized linear models. Their analysis depends heavily on the specific parametric noise models and it is unknown whether their algorithms can work with more general noise settings.

Under random classification noise, Balcan and Feldman [2013] proposes an algorithm which proceeds by estimating the distance between current linear separator and the optimal linear separator. However, their algorithm requires a suboptimal number of $\tilde{O}(\frac{d^2}{(1-2\eta)^2})$ labels. Their results also rely on the uniformity over the random classification noise, and it is shown in Awasthi et al. [2015] that this type of statistical query algorithms will fail in the heterogeneous noise setting (in particular the bounded noise setting).

Under bounded noise, the only known both label-efficient and computational-efficient algorithms are Awasthi et al. [2015, 2016]. Awasthi et al. [2015] uses a margin-based framework which only queries examples near decision boundary. To achieve computational efficiency, it adaptively chooses a sequence of hinge loss minimization problems to optimize instead of directly optimizing the 0-1 loss. Awasthi et al. [2015] only works when the noise probability upper bound η is extremely small ($\eta \leq 1.8 \times 10^{-6}$). Awasthi et al. [2016] improves over Awasthi et al. [2015] by adapting a polynomial regression procedure into the margin-based framework. Their algorithm works for any $\eta < 1/2$, but its label complexity is an unspecified high order polynomial with respect to d , which is far worse than the information-theoretic lower bound $\Omega(\frac{d}{(1-2\eta)^2} \log \frac{1}{\epsilon})$.

Under the more general adversarial noise setting, Awasthi et al. [2014] proposes a margin-based algorithm using a sequence of hinge loss minimization. Their algorithm can achieve an error of ϵ in polynomial time when $\nu = \Omega(\epsilon)$ (which is slightly better than our $\nu = \Omega(\frac{\epsilon}{\log d + \log \log \frac{1}{\epsilon}})$), but their algorithm requires $\tilde{O}(d^2)$ labels, which is suboptimal.

3 Definitions and Settings

We consider learning homogeneous linear separators under uniform distribution. The instance space \mathcal{X} is the unit sphere in \mathbb{R}^d , which we denote by $\mathbb{S}^{d-1} := \{x \in \mathbb{R}^d \mid \|x\| = 1\}$. We assume $d \geq 3$ throughout this paper. The label space $\mathcal{Y} = \{+1, -1\}$. We assume all data points (x, y) are drawn i.i.d. from an underlying distribution D over $\mathcal{X} \times \mathcal{Y}$. We denote by $D_{\mathcal{X}}$ the marginal of D over \mathcal{X} (which is uniform over \mathbb{S}^{d-1}), and $D_{Y|X}$ the conditional distribution of Y given X . Our algorithm is allowed to draw unlabeled examples $x \in \mathcal{X}$ from $D_{\mathcal{X}}$, and to query a labeling oracle \mathcal{O} for labels. Upon query x , \mathcal{O} returns a label y drawn from $D_{Y|X=x}$. The hypothesis class of interest is the set of homogeneous linear separators $\mathcal{H} := \{h_w(x) = \text{sign}(w \cdot x) \mid w \in \mathbb{S}^{d-1}\}$. For any hypothesis $h \in \mathcal{H}$, we define its error rate $\text{err}(h) = \mathbb{P}[h(X) \neq Y]$. Given a dataset $S = \{(X_1, Y_1), \dots, (X_m, Y_m)\} \in (\mathcal{X} \times \mathcal{Y})^m$, we define the empirical error rate of h over S as $\text{err}_S(h) = \frac{1}{m} \sum_{i=1}^m \mathbb{1}\{h(x_i) \neq y_i\}$.

For any two unit vectors v_1, v_2 , define $\theta(v_1, v_2) = \arccos(v_1 \cdot v_2)$ to be the angle between them. With some abuse of notations, we define $\theta(h_{v_1}, h_{v_2}) = \theta(v_1, v_2)$. It is easy to see that

$$|\text{err}(h_{v_1}) - \text{err}(h_{v_2})| \leq \mathbb{P}[h_{v_1}(x) \neq h_{v_2}(x)] = \frac{\theta(v_1, v_2)}{\pi}$$

Definition 1 (Bounded Noise). *We say the labeling oracle \mathcal{O} satisfies η -bounded noise condition for some $\eta \in [0, 1/2)$ with respect to u , if for any x , $\mathbb{P}[Y \neq \text{sign}(u \cdot x) \mid X = x] \leq \eta$.*

It is not hard to see that under η -bounded noise condition, h_u is the Bayes classifier. In addition, for any vector v , $|\text{err}(h_v) - \text{err}(h_u)| \geq \frac{1-2\eta}{\pi} \theta(v, u)$.

Definition 2 (Adversarial Noise). *We say the labeling oracle \mathcal{O} satisfies ν -adversarial noise condition for some $\nu \in [0, 1]$ with respect to u , if $\mathbb{P}[Y \neq \text{sign}(u \cdot X)] \leq \nu$.*

Given access to unlabeled examples drawn from $D_{\mathcal{X}}$ and a labeling oracle \mathcal{O} , our goal is to find a polynomial time algorithm that with probability at least $1 - \delta$, outputs a linear separator $h_v \in \mathcal{H}$ such that $\mathbb{P}[\text{sign}(v \cdot x) \neq \text{sign}(u \cdot x)] \leq \epsilon$ for some target accuracy ϵ and confidence δ^4 . The desired algorithm should make as few queries to the labeling oracle \mathcal{O} as possible.

⁴By triangle inequality, this type of guarantee can be readily converted to excess error guarantees, specifically, $\text{err}(h_v) - \text{err}(h_u) \leq \epsilon$.

We say an algorithm achieves a *label complexity* of $\Lambda(\epsilon, \delta)$, if for any $u \in \mathbb{S}^{d-1}$, with probability at least $1 - \delta$, it outputs a linear separator $h_v \in \mathcal{H}$ such that $\mathbb{P}[\text{sign}(v \cdot x) \neq \text{sign}(u \cdot x)] \leq \epsilon$, and requests at most $\Lambda(\epsilon, \delta)$ labels to oracle \mathcal{O} .

4 Main Algorithm

Our main algorithm, ACTIVE-PERCEPTRON (Algorithm 1), runs in epochs. At the beginning of each epoch k , it assumes a $\frac{\pi}{2^k}$ upper bound on $\theta(v_k, u)$, the angle between current iterate v_k and the underlying halfspace u . As we will see, this can be shown to hold with high probability inductively. Then, it calls procedure MODIFIED-PERCEPTRON (Algorithm 2) to find an improved estimate v_{k+1} , which can be shown to have an angle with u at most $\frac{\pi}{2^{k+1}}$ with high probability. The algorithm ends when a total of $\lceil \log_2 \frac{1}{\epsilon} \rceil$ epochs have passed.

We can assume without loss of generality that the angle between the initial halfspace v_0 and the underlying halfspace u is acute, that is, $\theta(v_0, u) \leq \frac{\pi}{2}$; Appendix F shows that such assumption can be removed with constant overhead in label and time complexity.

Algorithm 1 ACTIVE-PERCEPTRON

Input: Labeling oracle \mathcal{O} , initial halfspace v_0 , target error ϵ , confidence δ , noise upper bound η for bounded noise condition.

Output: learned halfspace \hat{v} .

- 1: Let $k_0 = \lceil \log_2 \frac{1}{\epsilon} \rceil$.
 - 2: **for** $k = 1, 2, \dots, k_0$ **do**
 - 3: $v_k \leftarrow \text{MODIFIED-PERCEPTRON}(\mathcal{O}, v_{k-1}, \frac{\pi}{2^k}, \frac{\delta}{k(k+1)}, \eta)$.
 - 4: **end for**
 - 5: **return** v_{k_0} .
-

The key component of the algorithm, MODIFIED-PERCEPTRON sequentially runs modified perceptron update rule [Motzkin and Schoenberg, 1954, Blum et al., 1998, Hampson and Kibler, 1999, Dasgupta et al., 2005] under a time-varying distribution $D_{\mathcal{X}}|_{R_t}$, where $R_t = \left\{x : \frac{b}{2} \leq w_t \cdot x \leq b\right\}$ is a band inside the unit sphere that contains examples that has an appropriate amount of projection along the current iterate w_t . The intuitive explanation of the region selection criterion has been presented in Subsection 1.2.

Algorithm 2 MODIFIED-PERCEPTRON

Input: Labeling oracle \mathcal{O} , initial halfspace w_0 , angle upper bound θ , confidence δ , noise upper bound η for bounded noise condition.

Output: Improved halfspace w_m .

- 1: Set parameters:
 - (i) Bounded Noise. Let $m = \lceil \frac{(3200\pi)^3 d}{(1-2\eta)^2} (\ln \frac{(3200\pi)^3 d}{(1-2\eta)^2} + \ln \frac{1}{\delta}) \rceil$, $b = \frac{\tilde{c}\theta(1-2\eta)}{\sqrt{d}}$, where $\tilde{c} = \frac{1}{2(600\pi)^2 \ln \frac{m^2}{\delta}}$.
 - (ii) Adversarial Noise. Let $m = \lceil (3200\pi)^3 d (\ln((3200\pi)^3 d) + \ln \frac{1}{\delta}) \rceil$, $b = \frac{\tilde{c}\theta}{\sqrt{d}}$, where $\tilde{c} = \frac{1}{2(600\pi)^2 \ln \frac{m^2}{\delta}}$.
 - 2: **for** $t = 0, 1, 2, \dots, m-1$ **do**
 - 3: Define region $R_t = \left\{x : \frac{b}{2} \leq w_t \cdot x \leq b\right\}$.
 - 4: Rejection sample $x_t \sim D_{\mathcal{X}}|_{R_t}$. Query \mathcal{O} for its label y_t .
 - 5: $w_{t+1} \leftarrow w_t - 2\mathbb{1}_{\{y_t w_t \cdot x_t < 0\}} \cdot (w_t \cdot x_t) \cdot x_t$.
 - 6: **end for**
 - 7: **return** w_m .
-

5 Performance Guarantees

5.1 A Lower Bound under Bounded Noise

We first present an information-theoretic lower bound of the label complexity in the bounded noise setting under uniform distribution. The proof of the theorem can be found at Appendix G.

Theorem 3. *For any $d > 4$, $0 \leq \eta < \frac{1}{2}$, $0 < \epsilon \leq \frac{1}{4\pi}$, $0 < \delta \leq \frac{1}{4}$, for any active learning algorithm \mathcal{A} , there is a $w^* \in \mathbb{S}^{d-1}$, and a labeling oracle \mathcal{O} that satisfies η -bounded noise condition with respect to w^* , such that if with probability at least $1 - \delta$, \mathcal{A} makes at most n queries of labels to \mathcal{O} and outputs $\hat{w} \in \mathbb{S}^{d-1}$ such that $\mathbb{P}[\text{sign}(\hat{w} \cdot x) \neq \text{sign}(w^* \cdot x)] \leq \epsilon$, then $n \geq \Omega\left(\frac{d \log \frac{1}{\epsilon}}{(1-2\eta)^2} + \frac{\eta \log \frac{1}{\delta}}{(1-2\eta)^2}\right)$.*

5.2 Bounded Noise

We establish Theorem 4 in the bounded noise setting. This gives the first computationally efficient active learning algorithm with near-optimal $\tilde{O}\left(\frac{d}{(1-2\eta)^2} \ln \frac{1}{\epsilon}\right)$ label complexity (see the lower bound theorem above) under the setting that $D_{\mathcal{X}}$ is uniform over the unit sphere and \mathcal{O} has bounded noise.

The proof of the theorem can be found at Appendix E. Appendix F shows that the assumption $\theta(v_0, u) \leq \frac{\pi}{2}$ can be removed with constant overhead in label complexity and time complexity.

Theorem 4 (ACTIVE-PERCEPTRON under Bounded Noise). *Suppose Algorithm 1 has inputs labeling oracle \mathcal{O} that satisfies η -bounded noise condition with respect to underlying halfspace u , initial halfspace v_0 such that $\theta(v_0, u) \leq \frac{\pi}{2}$, target error ϵ , confidence δ , then with probability $1 - \delta$:*

1. *The output halfspace v_k is such that $\mathbb{P}[\text{sign}(v_k \cdot x) \neq \text{sign}(u \cdot x)] \leq \epsilon$.*
2. *The total number of label queries is at most $O\left(\frac{d}{(1-2\eta)^2} \cdot \ln \frac{1}{\epsilon} \cdot \left(\ln \frac{d}{(1-2\eta)^2} + \ln \frac{1}{\delta} + \ln \ln \frac{1}{\epsilon}\right)\right)$.*
3. *The algorithm runs in time $O\left(\frac{d^2}{(1-2\eta)^3} \cdot \left(\ln \frac{d}{(1-2\eta)^2} + \ln \frac{1}{\delta} + \ln \ln \frac{1}{\epsilon}\right)^2 \cdot \frac{1}{\epsilon} \ln \frac{1}{\epsilon}\right)$.*

The theorem immediately follows from Lemma 1 below. The key ingredient of the lemma is a delicate analysis of the dynamics of the angles $\{\theta_t\}_{t=0}^m$, where $\theta_t := \theta(w_t, u)$ measures the closeness between the iterate w_t and the underlying halfspace u . Since x_t is randomly sampled and y_t is noisy, we are only able to show that θ_t decreases by a nonnegligible amount *in expectation*. To remedy this, we apply martingale concentration bounds to control the upper envelope of sequence $\{\theta_t\}_{t=0}^m$ carefully. The proof of the lemma can be found at Appendix D.

Lemma 1 (MODIFIED-PERCEPTRON under Bounded Noise). *Suppose Algorithm 2 has inputs labeling oracle \mathcal{O} that satisfies η -bounded noise condition with respect to underlying halfspace u , initial vector w_0 and angle upper bound θ such that $\theta(w_0, u) \leq \theta$, confidence δ , then with probability $1 - \delta$:*

1. *The output halfspace w_m is such that $\theta(w_m, u) \leq \frac{\theta}{2}$.*
2. *The number of label queries is at most $O\left(\frac{d}{(1-2\eta)^2} \left(\ln \frac{d}{(1-2\eta)^2} + \ln \frac{1}{\delta}\right)\right)$.*
3. *The algorithm runs in time $O\left(\frac{d^2}{(1-2\eta)^3} \cdot \left(\ln \frac{d}{(1-2\eta)^2} + \ln \frac{1}{\delta}\right)^2 \cdot \frac{1}{\theta}\right)$.*

5.3 Adversarial Noise

In the adversarial noise case, we show that ACTIVE-PERCEPTRON can tolerate a noise level of $\Omega(\frac{\epsilon}{\ln d + \ln \ln \frac{1}{\epsilon}})$. This is slightly weaker than $\Omega(\epsilon)$ shown in Awasthi et al. [2014]. However, we show that ACTIVE-PERCEPTRON has a near-optimal label complexity of $\tilde{O}(d \ln \frac{1}{\epsilon})$, which improves over the result of Awasthi et al. [2014] by a factor of d .

We present Theorem 5 below for the adversarial noise setting. The proof of the theorem can be found at Appendix E. Appendix F shows that the assumption $\theta(v_0, u) \leq \frac{\pi}{2}$ can be removed with constant overhead in label complexity and time complexity.

Theorem 5 (ACTIVE-PERCEPTRON under Adversarial Noise). *Suppose Algorithm 1 has inputs labeling oracle \mathcal{O} that satisfies ν -adversarial noise condition with respect to underlying halfspace u , initial halfspace v_0 such that $\theta(v_0, u) \leq \frac{\pi}{2}$, target error ϵ , confidence δ . Additionally $\nu \leq O(\frac{\epsilon}{\ln \frac{d}{\delta} + \ln \ln \frac{1}{\epsilon}})$. Then with probability $1 - \delta$:*

1. *The output halfspace v_k is such that $\mathbb{P}[\text{sign}(v_k \cdot x) \neq \text{sign}(u \cdot x)] \leq \epsilon$.*
2. *The total number of label queries is at most $O\left(d \cdot \ln \frac{1}{\epsilon} \cdot (\ln d + \ln \frac{1}{\delta} + \ln \ln \frac{1}{\epsilon})\right)$.*
3. *The algorithm runs in time $O\left(d^2 \cdot (\ln d + \ln \frac{1}{\delta} + \ln \ln \frac{1}{\epsilon})^2 \cdot \frac{1}{\epsilon} \ln \frac{1}{\epsilon}\right)$.*

The theorem immediately follows from Lemma 2 below, whose proof is similar to Lemma 1. The proof of the lemma can be found at Appendix D.

Lemma 2 (MODIFIED-PERCEPTRON under Adversarial Noise). *Suppose Algorithm 2 has inputs labeling oracle \mathcal{O} that satisfies ν -adversarial noise condition with respect to underlying halfspace u , initial vector w_0 and angle upper bound θ such that $\theta(w_0, u) \leq \theta$, confidence δ . Additionally $\nu \leq O(\frac{\epsilon}{\ln \frac{d}{\delta} + \ln \ln \frac{1}{\epsilon}})$. Then with probability $1 - \delta$:*

1. *The output halfspace w_m is such that $\theta(w_m, u) \leq \frac{\theta}{2}$.*
2. *The number of label queries is at most $O\left(d \cdot (\ln d + \ln \frac{1}{\delta})\right)$.*
3. *The algorithm runs in time $O\left(d^2 \cdot (\ln d + \ln \frac{1}{\delta})^2 \cdot \frac{1}{\epsilon}\right)$.*

6 Discussion and Open Problems

In this work, we propose a perceptron-based algorithm for efficient active learning of homogeneous linear separators under uniform distribution with noise. Under both bounded noise condition and adversarial noise condition, our algorithm achieves near-optimal label complexity. Our results significantly improves over the existing results in Awasthi et al. [2014, 2016].

Our analysis is performed under uniform unlabeled data distribution. However, it can be easily generalized to any spherical symmetrical distributions, for example, isotropic Gaussian distributions. It can also be generalized to distributions whose densities with respect to uniform distribution are bounded away from 0.

There are still many open problems in learning linear separators efficiently with respect to both time and label requirements:

1. Can perceptron-based algorithms, or other computationally efficient algorithms, be adapted to learn linear separator in the noisy setting, under more general data distributions (e.g. log-concave distributions) with optimal label complexity?

2. Are there any computationally efficient algorithms that can learn under weaker noise conditions, for example, Tsybakov noise condition [Tsybakov, 2004]?
3. Can we design efficient and label-optimal active learning algorithms without the knowledge of noise parameters?
4. Our algorithm is based on perceptron, a classical algorithm for online learning. Can one design other active learning algorithms through the lens of online learning? A representative example is Hanneke [2007b], where the algorithm simulates equivalence queries using label queries. Another interesting open question is, can one design a winnow-based algorithm [Littlestone, 1987] for attribute and computationally efficient active learning?

References

- Alekh Agarwal. Selective sampling algorithms for cost-sensitive multiclass prediction. *ICML (3)*, 28:1220–1228, 2013.
- Nir Ailon, Ron Begleiter, and Esther Ezra. Active learning using smooth relative regret approximations with applications. *Journal of Machine Learning Research*, 15(1):885–920, 2014.
- Martin Anthony and Peter L Bartlett. *Neural network learning: Theoretical foundations*. Cambridge University Press, 2009.
- Sanjeev Arora, László Babai, Jacques Stern, and Z Sweedyk. The hardness of approximate optima in lattices, codes, and systems of linear equations. In *Foundations of Computer Science, 1993. Proceedings., 34th Annual Symposium on*, pages 724–733. IEEE, 1993.
- Pranjal Awasthi, Maria Florina Balcan, and Philip M Long. The power of localization for efficiently learning linear separators with noise. In *Proceedings of the 46th Annual ACM Symposium on Theory of Computing*, pages 449–458. ACM, 2014.
- Pranjal Awasthi, Maria-Florina Balcan, Nika Haghtalab, and Ruth Uerner. Efficient learning of linear separators under bounded noise. In Peter Grünwald, Elad Hazan, and Satyen Kale, editors, *Proceedings of The 28th Conference on Learning Theory, COLT 2015, Paris, France, July 3-6, 2015*, volume 40 of *JMLR Proceedings*, pages 167–190. JMLR.org, 2015.
- Pranjal Awasthi, Maria-Florina Balcan, Nika Haghtalab, and Hongyang Zhang. Learning and 1-bit compressed sensing under asymmetric noise. In *Proceedings of The 28th Conference on Learning Theory, COLT 2016*, 2016.
- M.-F. Balcan and P. M. Long. Active and passive learning of linear separators under log-concave distributions. In *COLT*, 2013.
- M.-F. Balcan, A. Z. Broder, and T. Zhang. Margin based active learning. In *COLT*, 2007.
- M.-F. Balcan, A. Beygelzimer, and J. Langford. Agnostic active learning. *J. Comput. Syst. Sci.*, 75(1):78–89, 2009.
- Maria-Florina Balcan and Vitaly Feldman. Statistical active learning algorithms. In Christopher J. C. Burges, Léon Bottou, Zoubin Ghahramani, and Kilian Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States.*, pages 1295–1303, 2013.
- Maria-Florina Balcan, Steve Hanneke, and Jennifer Wortman Vaughan. The true sample complexity of active learning. *Machine learning*, 80(2-3):111–139, 2010.

- A. Beygelzimer, D. Hsu, J. Langford, and T. Zhang. Agnostic active learning without constraints. In *NIPS*, 2010.
- Alina Beygelzimer, Sanjoy Dasgupta, and John Langford. Importance weighted active learning. In *Twenty-Sixth International Conference on Machine Learning*, 2009.
- Avrim Blum, Alan M. Frieze, Ravi Kannan, and Santosh Vempala. A polynomial-time algorithm for learning noisy linear threshold functions. *Algorithmica*, 22(1/2):35–52, 1998. doi: 10.1007/PL00013833.
- David A. Cohn, Les E. Atlas, and Richard E. Ladner. Improving generalization with active learning. *Machine Learning*, 15(2):201–221, 1994.
- Koby Crammer, Yishay Mansour, Eyal Even-Dar, and Jennifer Wortman Vaughan. Regret minimization with concept drift. In *COLT*, pages 168–180. Citeseer, 2010.
- Amit Daniely. Complexity theoretic limitations on learning halfspaces. *arXiv preprint arXiv:1505.05800*, 2015.
- S. Dasgupta. Coarse sample complexity bounds for active learning. In *NIPS*, 2005.
- Sanjoy Dasgupta. Two faces of active learning. *Theoretical computer science*, 412(19):1767–1781, 2011.
- Sanjoy Dasgupta, Adam Tauman Kalai, and Claire Monteleoni. Analysis of perceptron-based active learning. In *Learning Theory, 18th Annual Conference on Learning Theory, COLT 2005, Bertinoro, Italy, June 27-30, 2005, Proceedings*, pages 249–263, 2005. doi: 10.1007/1150341517.
- Sanjoy Dasgupta, Daniel Hsu, and Claire Monteleoni. A general agnostic active learning algorithm. In *Advances in Neural Information Processing Systems 20*, 2007.
- Ofar Dekel, Claudio Gentile, and Karthik Sridharan. Selective sampling and active learning from single and multiple teachers. *Journal of Machine Learning Research*, 13(Sep):2655–2697, 2012.
- John Dunagan and Santosh Vempala. A simple polynomial-time rescaling algorithm for solving linear programs. In *Proceedings of the thirty-sixth annual ACM symposium on Theory of computing*, pages 315–320. ACM, 2004.
- Vitaly Feldman, Parikshit Gopalan, Subhash Khot, and Ashok Kumar Ponnuswami. New results for learning noisy parities and halfspaces. In *Foundations of Computer Science, 2006. FOCS’06. 47th Annual IEEE Symposium on*, pages 563–574. IEEE, 2006.
- Y. Freund, H. S. Seung, E. Shamir, and N. Tishby. Selective sampling using the query by committee algorithm. *Machine Learning*, 28(2-3):133–168, 1997.
- Venkatesan Guruswami and Prasad Raghavendra. Hardness of learning halfspaces with noise. *SIAM Journal on Computing*, 39(2):742–765, 2009.
- Steven Hampson and Dennis Kibler. Minimum generalization via reflection: A fast linear threshold learner. *Machine learning*, 37(1):51–73, 1999.
- S. Hanneke. A bound on the label complexity of agnostic active learning. In *ICML*, 2007a.
- S. Hanneke. *Theoretical Foundations of Active Learning*. PhD thesis, Carnegie Mellon University, 2009.
- Steve Hanneke. Teaching dimension and the complexity of active learning. In *International Conference on Computational Learning Theory*, pages 66–81. Springer, 2007b.
- Steve Hanneke. Rates of convergence in active learning. *The Annals of Statistics*, 39(1):333–361, 2011.

- Steve Hanneke, Varun Kanade, and Liu Yang. Learning with a drifting target concept. In *International Conference on Algorithmic Learning Theory*, pages 149–164. Springer, 2015.
- Steve Hanneke et al. Theory of disagreement-based active learning. *Foundations and Trends® in Machine Learning*, 7(2-3):131–309, 2014.
- D. Hsu. *Algorithms for Active Learning*. PhD thesis, UC San Diego, 2010.
- Tzu-Kuo Huang, Alekh Agarwal, Daniel Hsu, John Langford, and Robert E. Schapire. Efficient and parsimonious agnostic active learning. *CoRR*, abs/1506.08669, 2015.
- Adam Tauman Kalai, Adam R Klivans, Yishay Mansour, and Rocco A Servedio. Agnostically learning halfspaces. *SIAM Journal on Computing*, 37(6):1777–1805, 2008.
- Adam Klivans and Pravesh Kothari. Embedding Hard Learning Problems Into Gaussian Space. In Klaus Jansen, José D. P. Rolim, Nikhil R. Devanur, and Cristopher Moore, editors, *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques (APPROX/RANDOM 2014)*, volume 28 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 793–809, Dagstuhl, Germany, 2014. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik. ISBN 978-3-939897-74-3. doi: <http://dx.doi.org/10.4230/LIPIcs.APPROX-RANDOM.2014.793>.
- Adam R Klivans, Philip M Long, and Rocco A Servedio. Learning halfspaces with malicious noise. *Journal of Machine Learning Research*, 10(Dec):2715–2740, 2009.
- V. Koltchinskii. Rademacher complexities and bounding the excess risk in active learning. *JMLR*, 2010.
- Nick Littlestone. Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm. *Machine Learning*, 2(4):285–318, 1987. doi: 10.1007/BF00116827.
- Philip M Long. On the sample complexity of pac learning half-spaces against the uniform distribution. *IEEE Transactions on Neural Networks*, 6(6):1556–1559, 1995.
- Pascal Massart and Élodie Nédélec. Risk bounds for statistical learning. *The Annals of Statistics*, pages 2326–2366, 2006.
- Claire Monteleoni. Efficient algorithms for general active learning. In *International Conference on Computational Learning Theory*, pages 650–652. Springer, 2006.
- TS Motzkin and IJ Schoenberg. The relaxation method for linear inequalities. *Canadian Journal of Mathematics*, 6(3):393–404, 1954.
- Burr Settles. Active learning literature survey. *University of Wisconsin, Madison*, 52(55-66):11, 2010.
- A. B. Tsybakov. Optimal aggregation of classifiers in statistical learning. *Annals of Statistics*, 32:135–166, 2004.
- Liwei Wang. Smoothness, disagreement coefficient, and the label complexity of agnostic active learning. *Journal of Machine Learning Research*, 12(Jul):2269–2292, 2011.
- Chicheng Zhang and Kamalika Chaudhuri. Beyond disagreement-based agnostic active learning. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 442–450, 2014.

A Basic Lemmas

We collect a few useful facts in this section.

Lemma 3. *If $0 \leq x \leq 1 - \frac{1}{e}$, then for any $d \geq 1$, $(1 - \frac{x}{d})^{\frac{d}{2}} \geq e^{-x} \geq \frac{1}{2}$.*

Lemma 4. *Given $a \in (0, \pi)$, if $x \in [0, a]$, then $\frac{\sin a}{a}x \leq \sin x \leq x$.*

Lemma 5. *If $x \in [0, \pi]$, then $1 - \frac{x^2}{2} \leq \cos x \leq 1 - \frac{x^2}{5}$.*

Lemma 6. *Let $B(x, y) = \int_0^1 (1-t)^{x-1} t^{y-1} dt$ be the Beta function. Then $\frac{2}{\sqrt{d-1}} \leq B(\frac{1}{2}, \frac{d}{2}) \leq \frac{\pi}{\sqrt{d}}$.*

Lemma 7 (Marginal Density and Conditional Density). *If (x_1, x_2, \dots, x_d) is drawn from the uniform distribution over the unit sphere, then:*

1. (x_1, x_2) has density $p(z_1, z_2)$, where $p(z_1, z_2) = \frac{(1-z_1^2-z_2^2)^{\frac{d-4}{2}}}{\frac{2\pi}{d-2}}$.
2. Conditioned on $x_2 = b$, x_1 has density $p_b(z)$, where $p_b(z) = \frac{(1-b^2-z^2)^{\frac{d-4}{2}}}{(1-b^2)^{\frac{d-3}{2}} B(\frac{d-2}{2}, \frac{1}{2})}$.
3. x_1 has density $p(z)$, where $p(z) = \frac{(1-z^2)^{\frac{d-3}{2}}}{B(\frac{d-1}{2}, \frac{1}{2})}$.

Lemma 8 (Azuma's Inequality). *Let $\{Y_t\}_{t=1}^m$ be a bounded submartingale difference sequence, that is, $\mathbb{E}[Y_t | Y_1, \dots, Y_{t-1}] \geq 0$, and $|Y_t| \leq \sigma$. Then, with probability $1 - \delta$,*

$$\sum_{t=1}^m Y_t \geq -\sigma \sqrt{2m \ln \frac{1}{\delta}}$$

Lemma 9. *Suppose x is drawn uniformly from the unit sphere, and $b \leq \frac{1}{10\sqrt{d}}$. Then, $\mathbb{P} \left[x_1 \in \left[\frac{b}{2}, b \right] \right] \geq \frac{\sqrt{d}}{8\pi} b$.*

Proof.

$$\begin{aligned} & \mathbb{P} \left[x_1 \in \left[\frac{b}{2}, b \right] \right] \\ &= \frac{\int_{b/2}^b (1-t^2)^{\frac{d-3}{2}} dt}{B(\frac{d-1}{2}, \frac{1}{2})} \\ &\geq \frac{\frac{b}{2}(1-b^2)^{\frac{d-3}{2}}}{\frac{\pi}{\sqrt{d-1}}} \geq \frac{\sqrt{d}}{8\pi} b \end{aligned}$$

where the first equality is from item 3 of Lemma 7, giving the exact probability density function of x_1 , the first inequality is from that $(1-t^2)^{\frac{d-3}{2}} \geq (1-b^2)^{\frac{d-3}{2}}$ when $t \in [b/2, b]$, and Lemma 6 giving upper bound on $B(\frac{d-1}{2}, \frac{1}{2})$, and the second inequality is from Lemma 3 and that $d-1 \geq \frac{d}{2}$. \square

Lemma 10. *Suppose x is drawn uniformly from unit sphere restricted to the region $\{x : v \cdot x = \xi\}$, and u, v are unit vectors such that $\theta(u, v) = \theta \in [0, \frac{9}{10}\pi]$ and $0 \leq \xi \leq \frac{\theta}{4\sqrt{d}}$. Then,*

1. $\mathbb{E}[u \cdot x] \leq \xi$.
2. $\mathbb{E}[(u \cdot x)^2] \leq \frac{5\theta^2}{d}$.

$$3. \mathbb{E}[(u \cdot x) \mathbb{1}\{u \cdot x < 0\}] \leq \xi - \frac{\theta}{36\sqrt{d}}.$$

Proof. By spherical symmetry, without loss of generality, let $v = (0, 1, 0, \dots, 0)$, and $u = (\sin \theta, \cos \theta, 0, \dots, 0)$. Let $x = (x_1, \dots, x_d)$.

1.

$$\begin{aligned} & \mathbb{E}[u \cdot x] \\ = & \mathbb{E}[x_1 \sin \theta + x_2 \cos \theta | x_2 = \xi] \\ = & \mathbb{E}[x_1 | x_2 = \xi] \sin \theta + \xi \cos \theta \\ \leq & \xi \end{aligned}$$

where the first two equalities are by algebra, the inequality follows from $\cos \theta \leq 1$ and $\mathbb{E}[x_1 | x_2 = \xi] = 0$ since the conditional distribution of x_1 given $x_2 = \xi$ is symmetric around the origin.

2.

$$\begin{aligned} & \mathbb{E}[(u \cdot x)^2] \\ = & \mathbb{E}[(x_1 \sin \theta + x_2 \cos \theta)^2 | x_2 = \xi] \\ \leq & \mathbb{E}[2x_1^2 \sin^2 \theta + 2x_2^2 \cos^2 \theta | x_2 = \xi] \\ \leq & 2\mathbb{E}[x_1^2 | x_2 = \xi] \sin^2 \theta + 2\xi^2 \\ \leq & 2\theta^2 \frac{\int_{-1}^1 z^2 (1 - z^2)^{\frac{d-4}{2}} dz}{B(\frac{d-2}{2}, \frac{1}{2})} + 2\xi^2 \\ = & 2\theta^2 \frac{B(\frac{d-2}{2}, \frac{3}{2})}{B(\frac{d-2}{2}, \frac{1}{2})} + 2\xi^2 \\ \leq & \frac{5\theta^2}{d} \end{aligned}$$

where the first equality is by definition of u , the first inequality is from algebra that $(A + B)^2 \leq 2A^2 + 2B^2$, the second inequality is from that $|\cos \theta| \leq 1$, the third inequality is from item 2 of Lemma 7 and that $\sin \theta \leq \theta$, and the last inequality is from the fact that $\frac{B(\frac{d-2}{2}, \frac{3}{2})}{B(\frac{d-2}{2}, \frac{1}{2})} = \frac{1}{d-1} \leq \frac{2}{d}$, and $\xi^2 \leq \frac{\theta^2}{16d}$.

3.

$$\begin{aligned}
& \mathbb{E}[(u \cdot x) \mathbb{1}\{u \cdot x < 0\}] \\
&= \mathbb{E}[(x_1 \sin \theta + x_2 \cos \theta) \mathbb{1}\{x_1 < -\xi \cot \theta\} | x_2 = \xi] \\
&\leq \mathbb{E}[x_1 \mathbb{1}\{x_1 < -\xi \cot \theta\} | x_2 = \xi] \sin \theta + \xi \\
&= \xi + \sin \theta \int_{-\sqrt{1-\xi^2}}^{-\xi \cot \theta} \frac{(1 - \xi^2 - x_1^2)^{\frac{d-4}{2}} x_1}{(1 - \xi^2)^{\frac{d-3}{2}} \text{B}(\frac{d-2}{2}, \frac{1}{2})} dx_1 \\
&= \xi - \sin \theta \frac{\frac{2}{d-2} \left(1 - \left(\frac{\xi}{\sin \theta}\right)^2\right)^{\frac{d-2}{2}}}{(1 - \xi^2)^{\frac{d-3}{2}} \text{B}(\frac{d-2}{2}, \frac{1}{2})} \\
&\leq \xi - \sin \theta \frac{2}{\pi \sqrt{d-2}} \left(1 - \left(\frac{\xi}{\sin \theta}\right)^2\right)^{\frac{d-2}{2}} \\
&\leq \xi - \frac{\sin \theta}{\pi \sqrt{d}} \\
&\leq \xi - \frac{\theta}{36 \sqrt{d}}
\end{aligned}$$

where the first inequality is by algebra and $|\cos \theta| \leq 1$, the second equality is by item 2 of Lemma 7, the third equality is by integration, the second inequality is from $(1 - \xi^2)^{\frac{d-3}{2}} \leq 1$ and Lemma 6 that $\text{B}(\frac{d-2}{2}, \frac{1}{2}) \leq \frac{\pi}{\sqrt{d-2}}$, the third inequality follows by Lemma 3 that $\left(1 - \left(\frac{\xi}{\sin \theta}\right)^2\right)^{\frac{d-2}{2}} \geq \frac{1}{2}$, since $\xi \leq \frac{\theta}{4\sqrt{d}}$, and the last inequality follows from Lemma 4 that $\sin \theta \geq \frac{5\theta}{18\pi}$ when $\theta \in [0, \frac{9}{10}\pi]$ and algebra. \square

Lemma 11. Suppose Z_1, \dots, Z_n are iid Geometric(p) random variables. Then,

$$\mathbb{P}[Z_1 + \dots + Z_n > \frac{2n}{p}] \leq \exp(-\frac{n}{4})$$

Proof. Note that

$$\mathbb{P}[Z_1 + \dots + Z_n > \frac{2n}{p}] = \mathbb{P}[X_1 + \dots + X_{\lceil \frac{2n}{p} \rceil} < n]$$

where $X_1 \dots X_{\lceil \frac{2n}{p} \rceil}$ are iid Bernoulli(p) random variable. Therefore, by Chernoff bound, the above probability is at most $\exp(-\lceil \frac{2n}{p} \rceil \cdot p \cdot \frac{1}{8}) \leq \exp(-\frac{n}{4})$. \square

B Progress Measure under Bounded Noise

Lemma 12 (Progress in the Bounded Noise Model). Suppose $0 < \tilde{c} < \frac{1}{288}$, $b = \frac{\tilde{c}(1-2\eta)\theta}{\sqrt{d}}$, $\theta \leq \frac{27}{50}\pi$, and (x_t, y_t) is drawn from $D|_{R_t}$, where $R_t = \{(x, y) : x \cdot w_t \in [\frac{b}{2}, b]\}$ and the oracle \mathcal{O} satisfies η -bounded noise condition. If unit vector w_t has angle θ_t with u such that $\frac{1}{4}\theta \leq \theta_t \leq \frac{5}{3}\theta$, then update $w_{t+1} \leftarrow w_t - 2\mathbb{1}\{y_t \neq \text{sign}(w_t \cdot x_t)\} (w_t \cdot x_t) \cdot x_t$ has the following guarantee:

$$\mathbb{E}[\cos \theta_{t+1} - \cos \theta_t | \theta_t] \geq \frac{\tilde{c}}{100\pi} \frac{(1-2\eta)^2 \theta^2}{d}.$$

Proof. Define random variable $\xi = x_t \cdot w_t$. By the tower property of conditional expectation, $\mathbb{E} [\cos \theta_{t+1} - \cos \theta_t \mid \theta_t] = \mathbb{E} [\mathbb{E} [\cos \theta_{t+1} - \cos \theta_t \mid \theta_t, \xi] \mid \theta_t]$. Thus, it suffices to show

$$\mathbb{E} [\cos \theta_{t+1} - \cos \theta_t \mid \theta_t, \xi] \geq \frac{\tilde{c}}{100\pi} \frac{(1-2\eta)^2 \theta^2}{d}$$

for all $\theta_t \in [\frac{1}{4}\theta, \frac{5}{3}\theta]$ and $\xi \in [\frac{1}{2}b, b]$.

Recall that from Equation (2),

$$\cos \theta_{t+1} - \cos \theta_t = -2\mathbb{1} \{y_t \neq \text{sign}(w_t \cdot x_t)\} (w_t \cdot x_t) \cdot (u \cdot x_t).$$

We simplify $\mathbb{E} [\cos \theta_{t+1} - \cos \theta_t \mid \theta_t, \xi]$ as follows:

$$\begin{aligned} & \mathbb{E} [\cos \theta_{t+1} - \cos \theta_t \mid \theta_t, \xi] \\ &= \mathbb{E} [-2\xi u \cdot x_t \mathbb{1} \{y_t = -1\} \mid \theta_t, \xi] \\ &= \mathbb{E} [-2\xi u \cdot x_t (\mathbb{1} \{u \cdot x_t > 0, y_t = -1\} + \mathbb{1} \{u \cdot x_t < 0, y_t = -1\}) \mid \theta_t, \xi] \\ &\geq \mathbb{E} [-2\xi u \cdot x_t (\eta \mathbb{1} \{u \cdot x_t > 0\} + (1-\eta) \mathbb{1} \{u \cdot x_t < 0\}) \mid \theta_t, \xi] \\ &= \mathbb{E} [-2\xi u \cdot x_t (\eta + (1-2\eta) \mathbb{1} \{u \cdot x_t < 0\}) \mid \theta_t, \xi] \\ &= -2\xi \left(\eta \mathbb{E} [u \cdot x_t \mid \theta_t, \xi] + (1-2\eta) \mathbb{E} [u \cdot x_t \mathbb{1} \{u \cdot x_t < 0\} \mid \theta_t, \xi] \right) \end{aligned} \quad (3)$$

where the second equality is from algebra, the first inequality is from that $\mathbb{P}[y_t = -1 \mid u \cdot x_t > 0] \leq \eta$ and $\mathbb{P}[y_t = -1 \mid u \cdot x_t < 0] \geq 1 - \eta$, the last two equalities are from algebra.

By Lemma 10 and that $0 \leq \theta_t \leq \frac{5}{3}\theta \leq \frac{9}{10}\pi$, $\mathbb{E}[u \cdot x_t \mid \theta_t, \xi] \leq \xi$ and $\mathbb{E}[u \cdot x_t \mathbb{1} \{u \cdot x_t < 0\} \mid \theta_t, \xi] \leq \xi - \frac{\theta_t}{36\sqrt{d}}$.

Thus,

$$\begin{aligned} & \mathbb{E} [\cos \theta_{t+1} - \cos \theta_t \mid \theta_t, \xi] \\ &\geq -2\xi (\xi \eta + (\xi - \frac{\theta_t}{36\sqrt{d}})(1-2\eta)) \\ &\geq 2\xi (\frac{\theta_t}{36\sqrt{d}}(1-2\eta) - \xi) \\ &\geq b \frac{\theta_t}{72\sqrt{d}}(1-2\eta) \\ &\geq \frac{\tilde{c}}{100\pi} \frac{(1-2\eta)^2 \theta^2}{d} \end{aligned}$$

where the first and second inequalities are from algebra, the third inequality is from that $\xi \leq b \leq \frac{\theta(1-2\eta)}{288\sqrt{d}} \leq \frac{\theta_t(1-2\eta)}{72\sqrt{d}}$, and that $\xi \geq \frac{b}{2}$. the last inequality is by expanding $b = \frac{\tilde{c}(1-2\eta)\theta}{\sqrt{d}}$ and that $\theta_t \geq \frac{\theta}{4}$.

In conclusion, if $\frac{1}{4}\theta \leq \theta_t \leq \frac{5}{3}\theta$, then $\mathbb{E} [\cos \theta_{t+1} - \cos \theta_t \mid \theta_t, \xi] \geq \frac{\tilde{c}}{100\pi} \frac{(1-2\eta)^2 \theta^2}{d}$ for $\xi \in [\frac{b}{2}, b]$. The lemma follows. \square

C Progress Measure under Adversarial Noise

Lemma 13 (Progress in the Adversarial Noise Model). *Suppose $0 \leq \tilde{c} \leq \frac{1}{100\pi}$, $b = \frac{\tilde{c}\theta}{\sqrt{d}}$, $\theta \leq \frac{27}{50}\pi$, and (x_t, y_t) is drawn from distribution $D|_{R_t}$ where $R_t = \{(x, y) : x \cdot w_t \in [\frac{b}{2}, b]\}$. Meanwhile, the oracle \mathcal{O} satisfies ν -bounded noise condition where $\nu \leq \frac{\tilde{c}\theta}{192(200\pi)^2}$. If unit vector w_t has angle θ_t with u such that $\frac{1}{4}\theta \leq \theta_t \leq \frac{5}{3}\theta$, then update $w_{t+1} \leftarrow w_t - 2\mathbb{1} \{y_t \neq \text{sign}(w_t \cdot x_t)\} (w_t \cdot x_t) \cdot x_t$ has the following guarantee:*

$$\mathbb{E} [\cos \theta_{t+1} - \cos \theta_t \mid \theta_t] \geq \frac{\tilde{c}}{100\pi} \frac{\theta^2}{d}.$$

Proof. Define random variable $\xi = x_t \cdot w_t$.

Recall that from Equation (2),

$$\cos \theta_{t+1} - \cos \theta_t = -2\mathbb{1}\{y_t \neq \text{sign}(w_t \cdot x_t)\} (w_t \cdot x_t) \cdot (u \cdot x_t).$$

We expand $\mathbb{E} [\cos \theta_{t+1} - \cos \theta_t \mid \theta_t]$ as follows.

$$\begin{aligned} & \mathbb{E} [\cos \theta_{t+1} - \cos \theta_t \mid \theta_t] \\ = & \mathbb{E} [-2(w_t \cdot x_t)(u \cdot x_t)\mathbb{1}\{y_t = -1\} \mid \theta_t] \\ = & \mathbb{E} [-2(w_t \cdot x_t)(u \cdot x_t)\mathbb{1}\{u \cdot x_t < 0\} \mid \theta_t] \\ & + \mathbb{E} [2(w_t \cdot x_t)(u \cdot x_t)(\mathbb{1}\{y_t = +1, u \cdot x_t < 0\} - \mathbb{1}\{y_t = -1, u \cdot x_t > 0\}) \mid \theta_t] \end{aligned} \quad (4)$$

We bound the two terms separately. Firstly,

$$\begin{aligned} & \mathbb{E} [-2(w_t \cdot x_t)(u \cdot x_t)\mathbb{1}\{u \cdot x_t < 0\} \mid \theta_t] \\ \geq & -b\mathbb{E} [(u \cdot x_t)\mathbb{1}\{u \cdot x_t < 0\} \mid \theta_t] \\ = & -b\mathbb{E} [\mathbb{E} [(u \cdot x_t)\mathbb{1}\{u \cdot x_t < 0\} \mid \theta_t, b] \mid \theta_t] \\ \geq & b\left(\frac{\theta_t}{36\sqrt{d}} - b\right) \end{aligned} \quad (5)$$

where the first inequality is from that $-(u \cdot x_t)\mathbb{1}\{u \cdot x_t < 0\} \geq 0$ and $w_t \cdot x_t \geq \frac{b}{2}$, the equality is from the tower property of conditional expectation, the second inequality is from Lemma 10.

Secondly,

$$\begin{aligned} & \left| \mathbb{E} [2(w_t \cdot x_t)(u \cdot x_t)(\mathbb{1}\{y_t = +1, u \cdot x_t < 0\} - \mathbb{1}\{y_t = -1, u \cdot x_t > 0\}) \mid \theta_t] \right| \\ \leq & 2b\mathbb{E} [|u \cdot x_t|\mathbb{1}\{y_t \neq \text{sign}(u \cdot x_t)\} \mid \theta_t] \\ \leq & 2b\sqrt{\mathbb{E} [\mathbb{1}\{y_t \neq \text{sign}(u \cdot x_t)\} \mid \theta_t] \cdot \mathbb{E} [(u \cdot x_t)^2 \mid \theta_t]} \\ = & 2b\sqrt{\mathbb{P}[y_t \neq \text{sign}(u \cdot x_t) \mid \theta_t] \mathbb{E} [\mathbb{E} [(u \cdot x_t)^2 \mid \theta_t, \xi] \mid \theta_t]} \end{aligned} \quad (6)$$

where the first inequality is from that $|\mathbb{E}[X]| \leq \mathbb{E}[|X|]$, and $w_t \cdot x_t \leq b$, the second inequality is from Cauchy-Schwarz, the third equality is by algebra.

Now we look at the two terms inside the square root. First,

$$\begin{aligned} & \mathbb{P}[y_t \neq \text{sign}(u \cdot x_t) \mid \theta_t] \\ = & \mathbb{P}_{x \sim D|_{R_t}} [y \neq \text{sign}(u \cdot x)] \\ \leq & \frac{\mathbb{P}_{(x,y) \sim D} [y \neq \text{sign}(u \cdot x)]}{\mathbb{P}_{x \sim D} [x_1 \in [b/2, b]]} \\ \leq & \frac{8\pi\nu}{\tilde{c}\theta} \\ \leq & \frac{1}{16(200\pi)^2} \end{aligned}$$

where the first inequality is from that $\mathbb{P}[A|B] \leq \frac{\mathbb{P}[A]}{\mathbb{P}[B]}$, the second inequality is from Lemma 9 that $\mathbb{P}_{x \sim D} [x_1 \in [b/2, b]] \geq \frac{\sqrt{d}b}{8\pi} = \frac{\tilde{c}\theta}{8\pi}$, and the last inequality is by our assumption on ν .

Second, fix $\xi \in [\frac{b}{2}, b]$, $\xi \leq b \leq \frac{\theta_t}{4\sqrt{d}}$. Item 2 of Lemma 10 implies that $\mathbb{E}[(u \cdot x_t)^2 \mid \theta_t, \xi] \leq \frac{5\theta_t^2}{d}$. By the tower property of conditional expectation, $\mathbb{E}[(u \cdot x_t)^2 \mid \theta_t] \leq \frac{5\theta_t^2}{d}$. Continuing Equation (6), we get

$$\left| \mathbb{E} [2(w_t \cdot x_t)(u \cdot x_t)(\mathbb{1}\{y_t = +1, u \cdot x_t < 0\} - \mathbb{1}\{y_t = -1, u \cdot x_t > 0\}) \mid \theta_t] \right| \leq b \frac{\theta_t}{100\pi\sqrt{d}}. \quad (7)$$

Continuing Equation (4), we have

$$\begin{aligned} & \mathbb{E} [\cos \theta_{t+1} - \cos \theta_t \mid \theta_t] \\ & \geq b \left(\frac{\theta_t}{36\sqrt{d}} - \frac{\theta_t}{100\pi\sqrt{d}} - b \right) \\ & \geq b \frac{\theta_t}{25\pi\sqrt{d}} \geq \frac{\tilde{c}}{100\pi} \frac{\theta^2}{d} \end{aligned}$$

where the first inequality is from Equations (5) and (7), the second inequality is from algebra and that $b \leq \frac{\theta_t}{100\pi\sqrt{d}}$, the third inequality is by expanding $b = \frac{\tilde{c}\theta}{\sqrt{d}}$ and $\theta_t \geq \frac{\theta}{4}$. \square

D Performance Guarantees of Modified-Perceptron

First we need a technical bound on the difference between $\cos \theta_{t+1}$ and $\cos \theta_t$ that holds with probability 1.

Lemma 14. *Suppose $0 < \tilde{c} < 1$, $b = \frac{\tilde{c}(1-2\eta)\theta}{\sqrt{d}} \leq 1$, and (x_t, y_t) is drawn from distribution $D|_{R_t}$ where $R_t = \{(x, y) : x \cdot w_t \in [\frac{b}{2}, b]\}$. If unit vector w_t has angle $\theta_t \leq \frac{5}{3}\theta$, then update $w_{t+1} \leftarrow w_t - 2\mathbb{1}\{y_t \neq \text{sign}(w_t \cdot x_t)\}(w_t \cdot x_t) \cdot x_t$ has the following guarantee: $|\cos \theta_{t+1} - \cos \theta_t| \leq \frac{16\tilde{c}(1-2\eta)\theta^2}{3\sqrt{d}}$.*

Proof. Recall that from Equation (2),

$$\cos \theta_{t+1} - \cos \theta_t = -2\mathbb{1}\{y_t \neq \text{sign}(w_t \cdot x_t)\}(w_t \cdot x_t) \cdot (u \cdot x_t).$$

Firstly, note $|\cos \theta_{t+1} - \cos \theta_t| \leq 2|w_t \cdot x_t||u \cdot x_t| \leq 2b|u \cdot x_t|$.

Observe that

$$\begin{aligned} & |u \cdot x_t| \\ & \leq |w_t \cdot x_t| + |(u - w_t) \cdot x_t| \\ & \leq b + 2 \sin \frac{\theta_t}{2} \\ & \leq b + \theta_t \end{aligned}$$

Thus, we have $|\cos \theta_{t+1} - \cos \theta_t| \leq 2b(b + \theta_t) = \frac{2\tilde{c}^2(1-2\eta)^2\theta^2}{d} + \frac{2\tilde{c}(1-2\eta)\theta\theta_t}{\sqrt{d}} \leq \frac{16\tilde{c}(1-2\eta)\theta^2}{3\sqrt{d}}$. \square

Lemma 15. *Suppose:*

1. Initial unit vector w_0 has angle $\theta_0 = \theta(w_0, u) \leq \theta \leq \frac{27}{50}\pi$ with u ;
2. Integer $m = \lceil \frac{(3200\pi)^3 d}{(1-2\eta)^2} (\ln \frac{(3200\pi)^3 d}{(1-2\eta)^2} + \ln \frac{1}{\delta}) \rceil$ and $\tilde{c} = \frac{1}{2(600\pi)^2 \ln \frac{m^2}{8}}$;
3. For all t , if $\frac{1}{4}\theta \leq \theta_t \leq \frac{5}{3}\theta$, then $\mathbb{E}[\cos \theta_{t+1} - \cos \theta_t \mid \theta_t] \geq \frac{\tilde{c}}{100\pi} \frac{(1-2\eta)^2 \theta^2}{d}$;
4. For all t , if $\theta_t \leq \frac{5}{3}\theta$, then $|\cos \theta_{t+1} - \cos \theta_t| \leq \frac{16\tilde{c}(1-2\eta)\theta^2}{3\sqrt{d}}$ holds with probability 1.

Then with probability $1 - \delta$, after m iterations in MODIFIED-PERCEPTRON,

1. The number of label queries to oracle \mathcal{O} is at most $m = O(\frac{d}{(1-2\eta)^2} \log \frac{d}{\delta(1-2\eta)^2})$.
2. The running time of the algorithm is at most $T = O(\frac{d^2}{(1-2\eta)^3} \log^2 \frac{d}{\delta(1-2\eta)^2} \frac{1}{\theta})$.
3. The output w_m is such that $\theta_m \leq \frac{1}{2}\theta$;

Proof. First, the number of label queries is m , which is

$$\lceil \frac{(3200\pi)^3 d}{(1-2\eta)^2} (\ln \frac{(3200\pi)^3 d}{(1-2\eta)^2} + \ln \frac{1}{\delta}) \rceil = O\left(\frac{d}{(1-2\eta)^2} \log \frac{d}{\delta(1-2\eta)^2}\right).$$

Second, we analyze the time complexity of the algorithm. At each iteration $t \in [0, m]$, it takes Z_t trials to hit an example in $[\frac{b}{2}, b]$, where Z_t is a Geometric(p) random variable with $p = \mathbb{P}_{x \sim D_X}[w_t \cdot x \in [\frac{b}{2}, b]]$. From Lemma 9, $p \geq \frac{\sqrt{d}}{8\pi} b = \frac{\tilde{c}(1-2\eta)\theta}{8\pi} = \Omega(\frac{(1-2\eta)\theta}{\ln \frac{d}{\delta(1-2\eta)^2}})$.

Define event

$$E_1 := \left\{ Z_1 + \dots + Z_m \leq \frac{2m}{p} \right\}$$

From Lemma 11 and the choice of m , $\mathbb{P}[E_1] \geq 1 - \frac{\delta}{2}$. Thus, on event E_1 , the total number of rejection sampling trials is at most $O(\frac{d}{(1-2\eta)^3} \log^2 \frac{d}{\delta(1-2\eta)^2} \frac{1}{\theta})$.

Lastly, we prove the upper bound of the angle θ_m . Define random variable D_t as:

$$D_t := \left(\cos \theta_{t+1} - \cos \theta_t - \frac{\tilde{c}}{100\pi} \frac{(1-2\eta)^2 \theta^2}{d} \right) \mathbb{1} \left\{ \frac{1}{4}\theta \leq \theta_t \leq \frac{5}{3}\theta \right\}$$

Note that $\mathbb{E}[D_t | \theta_t] \geq 0$ and from Lemma 14, $|D_t| \leq |\cos \theta_{t+1} - \cos \theta_t| + \frac{\tilde{c}}{100\pi} \frac{(1-2\eta)^2 \theta^2}{d} \leq \frac{6\tilde{c}(1-2\eta)\theta^2}{\sqrt{d}}$. Therefore, $\{D_t\}$ is a bounded submartingale difference sequence. By Azuma's Inequality (see Lemma 8) and union bound, define event

$$E_2 = \left\{ \text{for all } 0 \leq t_1 \leq t_2 \leq m, \sum_{s=t_1}^{t_2-1} D_s \geq -\frac{6\tilde{c}(1-2\eta)\theta^2}{\sqrt{d}} \sqrt{2(t_2 - t_1) \ln \frac{2m^2}{\delta}} \right\}$$

Then $\mathbb{P}(E_2) \geq 1 - \frac{\delta}{2}$.

Now we condition on event E_2 . We break the subsequent analysis into two parts: (1) Show there exists some t such that θ_t goes below $\frac{1}{4}\theta$. (2) Show that, afterwards, θ_t must stay below $\frac{1}{2}\theta$.

1. First, it can be checked by algebra that $m \geq \frac{200\pi d}{(1-2\eta)^2 \tilde{c}}$. We show the following claim.

Claim 1. *There exists some $t \in [0, m]$, such that $\theta_t < \frac{1}{4}\theta$.*

Proof. We first show that it is impossible for all $t \in [0, m]$ such that $\theta_t \in [\frac{1}{4}\theta, \frac{5}{3}\theta]$. To this end, assume this holds for the sake of contradiction. In this case, for all $t \in [0, m]$, $D_t = \cos \theta_{t+1} - \cos \theta_t - \frac{\tilde{c}}{100\pi} \frac{(1-2\eta)^2 \theta^2}{d}$. Therefore,

$$\begin{aligned} & \cos \theta_m - \cos \theta_0 \\ &= \sum_{s=0}^{m-1} D_s + \frac{\tilde{c}}{100\pi} \frac{(1-2\eta)^2 \theta^2}{d} m \\ &\geq \frac{\tilde{c}}{100\pi} \frac{(1-2\eta)^2 \theta^2}{d} m - \frac{6\tilde{c}(1-2\eta)\theta^2}{\sqrt{d}} \sqrt{2m \ln \frac{m^2}{\delta}} \\ &\geq \frac{\theta^2}{100\pi} \left[\frac{\tilde{c}(1-2\eta)^2 m}{d} - \sqrt{\frac{\tilde{c}(1-2\eta)^2 m}{d}} \right] \\ &\geq \theta^2 \end{aligned}$$

where the first inequality is from the definition of event E_1 , the second inequality is from that $\tilde{c} = \frac{1}{2(600\pi)^2 \ln \frac{m^2}{\delta}}$, the third inequality is from that $\frac{\tilde{c}(1-2\eta)^2 m}{d} \geq 200\pi$.

Since $\cos \theta_0 \geq \cos \theta \geq 1 - \frac{1}{2}\theta^2$, this gives that $\cos \theta_m \geq 1 + \frac{1}{2}\theta^2 > 1$, contradiction.

Next, define $\tau := \min \left\{ t \geq 0 : \theta_t \notin \left[\frac{1}{4}\theta, \frac{5}{3}\theta \right] \right\}$. We now know that $\tau \leq m$. It suffices to show that $\theta_\tau < \frac{1}{4}\theta$, that is, the first time θ_t goes outside $\left[\frac{1}{4}\theta, \frac{5}{3}\theta \right]$, it must be crossing the left boundary.

By definition of τ , for all $0 \leq t \leq \tau - 1$, $\theta_t \in \left[\frac{1}{4}\theta, \frac{5}{3}\theta \right]$. Thus,

$$\begin{aligned} & \cos \theta_\tau - \cos \theta_0 \\ &= \sum_{t=0}^{\tau-1} D_t + \frac{\tilde{c}}{100\pi} \frac{(1-2\eta)^2 \theta^2}{d} \tau \\ &\geq \frac{\tilde{c}}{100\pi} \frac{(1-2\eta)^2 \theta^2}{d} \tau - \frac{6\tilde{c}(1-2\eta)\theta^2}{\sqrt{d}} \sqrt{\tau \ln \frac{m^2}{\delta}} \\ &\geq -900\pi \ln \frac{m^2}{\delta} \tilde{c} \theta^2 \geq -\frac{1}{75} \theta^2 \end{aligned} \tag{8}$$

where the first inequality is by the definition of E_2 , and the second inequality is by minimization over $\tau \in [0, m]$, the last inequality is from the definition of \tilde{c} .

Now, if $\theta_\tau \geq \frac{5}{3}\theta$, then

$$\begin{aligned} \cos \theta_\tau - \cos \theta_0 &\leq \cos \frac{5}{3}\theta - \cos \theta \\ &\leq 1 - \frac{1}{5} \left(\frac{5}{3} \right)^2 \theta^2 - 1 + \frac{1}{2} \theta^2 \\ &< -\frac{1}{75} \theta^2 \end{aligned}$$

where the first inequality follows from $\theta_\tau \geq \frac{5}{3}\theta$ and $\theta_0 \leq \theta$, and the second inequality follows from Lemma 5. This contradicts with Inequality (8).

Therefore, $\theta_\tau < \frac{5}{3}\theta$. Since $\theta_\tau \notin \left[\frac{1}{4}\theta, \frac{5}{3}\theta \right]$, $\theta_\tau < \frac{1}{4}\theta$. \square

2. We now show the following claim to conclude the proof.

Claim 2. $\theta_m \leq \frac{1}{2}\theta$.

Proof. Define $\sigma = \max \{ t \in [0, m] : \theta_t < \frac{1}{4}\theta \}$. We have shown such σ exists. We show afterwards, θ_t will not exceed $\frac{1}{2}\theta$. Assume for contradiction that for some $t > \sigma$, $\theta_t > \frac{1}{2}\theta$.

Now define $\gamma := \min \{ t > \sigma : \theta_t > \frac{1}{2}\theta \}$. We know by definition of σ and γ , for all $t \in [\sigma + 1, \gamma - 1]$, $\theta_t \in \left[\frac{1}{4}\theta, \frac{1}{2}\theta \right]$. Thus,

$$\begin{aligned} & \cos \theta_\gamma - \cos \theta_{\sigma+1} \\ &= \sum_{t=\sigma+1}^{\gamma-1} D_t + \frac{\tilde{c}}{100\pi} \frac{(1-2\eta)^2 \theta^2}{d} (\gamma - \sigma - 1) \\ &\geq \frac{\tilde{c}}{100\pi} \frac{(1-2\eta)^2 \theta^2}{d} (\gamma - \sigma - 1) - \frac{6\tilde{c}(1-2\eta)\theta^2}{\sqrt{d}} \sqrt{(\gamma - \sigma - 1) \ln \frac{m^2}{\delta}} \\ &\geq -900\pi \ln \frac{m^2}{\delta} \tilde{c} \geq -\frac{1}{75} \theta^2 \end{aligned} \tag{9}$$

where the first inequality is by the definition of E_2 , and the second inequality is by minimization over $\gamma - \sigma - 1 \in [0, m]$, the last inequality is from the definition of \tilde{c} .

On the other hand, $\theta_\gamma > \frac{1}{2}\theta$ and $\theta_\sigma < \frac{1}{4}\theta$. We have

$$\begin{aligned} \cos \theta_\gamma - \cos \theta_{\sigma+1} &\leq \cos \theta_\gamma - \cos \theta_\sigma + \frac{6\tilde{c}(1-2\eta)\theta^2}{\sqrt{d}} \\ &\leq \cos \frac{\theta}{2} - \cos \frac{\theta}{4} + \frac{6\tilde{c}(1-2\eta)\theta^2}{\sqrt{d}} \\ &\leq 1 - \frac{1}{20}\theta^2 - 1 + \frac{1}{32}\theta^2 + \frac{6\tilde{c}(1-2\eta)\theta^2}{\sqrt{d}} \\ &< -\frac{1}{75}\theta^2 \end{aligned}$$

where the first inequality follows from Lemma 14, the third follows from Lemma 5, and the last follows from algebra. This contradicts with Inequality (9). \square

Thus, on event $E := E_1 \cap E_2$, item 1,2,3 hold simultaneously. By union bound, this holds with probability $1 - \delta$. \square

Proof of Lemma 1. This is an immediate consequence of Lemmas 12 and 15. \square

Proof of Lemma 2. This is an immediate consequence of Lemmas 13 and 15. \square

E Proofs of Theorems 4 and 5

Proof of Theorem 4. From Lemma 1, we know that for every k , there is an event E_k such that $\mathbb{P}(E_k) \geq 1 - \frac{\delta}{k(k+1)}$, and on event E_k , items 1,2,3 of Lemma 1 holds for input $w_0 = v_k$, output $w_m = v_{k+1}$, $\theta = \frac{\pi}{2^k}$.

Define event $E = \cup_{k=1}^{k_0} E_k$. By union bound, $\mathbb{P}(E) \geq 1 - \delta$. We henceforth condition on event E happening.

1. By induction, the final output v_{k_0} has the property that $\theta(v_k, u) \leq 2^{-k_0}\pi \leq \epsilon\pi$, implying that $\mathbb{P}[\text{sign}(v_k \cdot x) \neq \text{sign}(u \cdot x)] \leq \epsilon$.
2. Define the number of label queries to oracle \mathcal{O} at iteration k as m_k . From Lemma 15, m_k is at most $O\left(\frac{d}{(1-2\eta)^2} \left(\ln \frac{d}{(1-2\eta)^2} + \ln \frac{k}{\delta}\right)\right)$. Thus, the total number of label queries to oracle \mathcal{O} is $\sum_{k=1}^{k_0} m_k$, which is at most

$$k_0 \cdot m_{k_0} = O\left(k_0 \cdot \frac{d}{(1-2\eta)^2} \left(\ln \frac{d}{(1-2\eta)^2} + \ln \frac{k_0}{\delta}\right)\right).$$

Item 2 is proved by noting $k_0 \leq \log \frac{1}{\epsilon} + 1$.

3. Define the running time of MODIFIED-PERCEPTRON at iteration k as T_k . From Lemma 15, T_k is at most $O\left(\frac{d^2}{(1-2\eta)^3} \cdot \left(\ln \frac{d}{(1-2\eta)^2} + \ln \frac{k}{\delta}\right)^2 \cdot \frac{1}{\epsilon}\right)$. Thus, the total running time is at most $\sum_{k=1}^{k_0} T_k$, which is at most

$$k_0 T_{k_0} = O\left(k_0 \cdot \frac{d^2}{(1-2\eta)^3} \cdot \left(\ln \frac{d}{(1-2\eta)^2} + \ln \frac{k_0}{\delta}\right)^2 \cdot \frac{1}{\epsilon}\right).$$

Item 3 is proved by noting $k_0 \leq \log \frac{1}{\epsilon} + 1$. \square

Proof of Theorem 5. The proof is identical to the proof of Theorem 5 taking $\eta = 0$. \square

F Technical Details on Acute Initialization

We argue in this section that the angle between the initial vector v_0 and the optimal hypothesis u can be assumed to be acute without loss of generality in the bounded noise setting. A similar result has been shown by Awasthi et al. [2014], Appendix B for the adversarial noise setting. To this end, we construct an Algorithm 3 an initialization procedure. It runs ACTIVE-PERCEPTRON twice, taking a vector v_0 and its negation $-v_0$ as initializers. Then it performs hypothesis testing using $\tilde{O}(\frac{1}{(1-2\eta)^2})$ labeled examples to find out a halfspace which has angle at most $\frac{\pi}{4}$ with u .

Algorithm 3 Master Algorithm

Input: Labeling oracle \mathcal{O} , confidence δ , noise upper bound η for bounded noise condition.

Output: a halfspace \hat{v} such that $\theta(\hat{v}, v^*) \leq \frac{\pi}{4}$.

```

1:  $v_0 \leftarrow$  an arbitrary vector in  $\mathbb{S}^{d-1}$ .
2:  $v_+ \leftarrow \text{ACTIVE-PERCEPTRON}(\mathcal{O}, v_0, \frac{(1-2\eta)}{16}, \frac{\delta}{3}, \eta)$ .
3:  $v_- \leftarrow \text{ACTIVE-PERCEPTRON}(\mathcal{O}, -v_0, \frac{(1-2\eta)}{16}, \frac{\delta}{3}, \eta)$ .
4: Define region  $R := \{x : \text{sign}(v_+ \cdot x) \neq \text{sign}(v_- \cdot x)\}$ .
5:  $S \leftarrow$  Draw  $\frac{8}{(1-2\eta)^2} \ln \frac{6}{\delta}$  iid samples from  $D|_R$  and query their labels.
6: if  $\text{err}_S(h_{v_+}) \leq \text{err}_S(h_{v_-})$  then
7:   return  $v_+$ 
8: else
9:   return  $v_-$ 
10: end if
```

We show Algorithm 3 learns the target halfspace unconditionally with a constant overhead of label complexity and time complexity.

Theorem 6. Suppose Algorithm 3 has inputs labeling oracle \mathcal{O} that satisfies η -bounded noise condition with respect to u , confidence δ . Then, with probability $1 - \delta$, the output \hat{v} is such that $\theta(\hat{v}, u) \leq \frac{\pi}{4}$. Furthermore, the total number of label queries to oracle \mathcal{O} is at most $\tilde{O}\left(\frac{d}{(1-2\eta)^2}\right)$ and the algorithm runs in time $\tilde{O}\left(\frac{d^2}{(1-2\eta)^3}\right)$.

Proof. Note that one of $\theta(v_0, u)$, $\theta(-v_0, u)$ is at most $\frac{\pi}{2}$. From Theorem 4 and union bound, we know that with probability $1 - \frac{2\delta}{3}$, either $\theta(v_+, u) \leq \frac{(1-2\eta)\pi}{16}$, or $\theta(v_-, u) \leq \frac{(1-2\eta)\pi}{16}$.

Suppose without loss of generality, $\theta(v_+, u) \leq \frac{(1-2\eta)\pi}{16}$. We consider two cases.

Case 1: $\theta(v_+, v_-) \leq \pi/8$. By triangle inequality, $\theta(v_-, u) \leq \theta(v_+, u) + \theta(v_+, v_-) \leq \pi/4$. In this case, $\theta(v_+, u) \leq \frac{\pi}{4}$ and $\theta(v_-, u) \leq \frac{\pi}{4}$ holds simultaneously. Therefore, the returned vector \hat{v} satisfies $\theta(\hat{v}, u) \leq \frac{\pi}{4}$.

Case 2: $\theta(v_+, v_-) > \pi/8$. In this case, $\mathbb{P}[x \in R] \geq 1/8$, thus,

$$\mathbb{P}_R[\text{sign}(v_+ \cdot x) \neq \text{sign}(u \cdot x)] \leq \frac{\mathbb{P}[\text{sign}(v_+ \cdot x) \neq \text{sign}(u \cdot x)]}{\mathbb{P}[x \in R]} \leq \frac{1-2\eta}{8} = \frac{1}{4}\left(\frac{1}{2} - \eta\right).$$

Meanwhile, $\mathbb{P}_R[\text{sign}(v_+ \cdot x) \neq y] \leq \eta \mathbb{P}_R[\text{sign}(v_+ \cdot x) = \text{sign}(u \cdot x)] + \mathbb{P}_R[\text{sign}(v_+ \cdot x) \neq \text{sign}(u \cdot x)]$. Therefore,

$$\begin{aligned}
& \frac{1}{2} - \mathbb{P}_R[\text{sign}(v_+ \cdot x) \neq y] \\
& \geq \left(\frac{1}{2} - \eta\right) \mathbb{P}_R[\text{sign}(v_+ \cdot x) = \text{sign}(u \cdot x)] - \frac{1}{2} \mathbb{P}_R[\text{sign}(v_+ \cdot x) \neq \text{sign}(u \cdot x)] \\
& \geq \left(\frac{1}{2} - \eta\right) \cdot \frac{1}{2} - \left(\frac{1}{2} - \eta\right) \cdot \frac{1}{4} \\
& \geq \frac{1}{4}\left(\frac{1}{2} - \eta\right)
\end{aligned}$$

Since v_+ disagrees with v_- everywhere on R , $\mathbb{P}_R[\text{sign}(v_+ \cdot x) \neq y] + \mathbb{P}_R[\text{sign}(v_- \cdot x) \neq y] = 1$. Thus, $\text{err}_{D|R}(h_{v_+}) \leq \frac{1}{2} - (\frac{1}{2} - \eta)\frac{1}{4}$ and $\text{err}_{D|R}(h_{v_-}) \geq \frac{1}{2} + (\frac{1}{2} - \eta)\frac{1}{4}$. Therefore, by Hoeffding's Inequality, with probability $1 - \delta/3$,

$$\text{err}_S(v_+) < \frac{1}{2} < \text{err}_S(v_-)$$

therefore v_+ will be selected for \hat{v} . This gives that $\theta(\hat{v}, u) \leq \pi/4$.

To conclude, by union bound, we have shown that with probability $1 - \delta$, $\theta(\hat{v}, u) \leq \frac{\pi}{4}$. The time complexity and label complexity follows immediately from Theorem 4. \square

G Proof of the Lower Bound

In this section, we give the proof of Theorem 3. It follows from two key lemmas, Lemma 20 and Lemma 21. First we start with some additional definitions.

Definition 3. Let \mathbb{P}, \mathbb{Q} be two probability measures on a common measurable space and \mathbb{P} is absolutely continuous with respect to \mathbb{Q} .

- The KL-divergence between \mathbb{P} and \mathbb{Q} is defined as $D_{KL}(\mathbb{P}, \mathbb{Q}) = \mathbb{E}_{X \sim \mathbb{P}} \ln \frac{\mathbb{P}(X)}{\mathbb{Q}(X)}$.
- We define $d_{KL}(p, q) = D_{KL}(\mathbb{P}, \mathbb{Q})$, where \mathbb{P}, \mathbb{Q} are distributions of a Bernoulli(p) and a Bernoulli(q) random variables respectively.
- For random variables X, Y, Z , define the mutual information between X and Y under \mathbb{P} as $I(X; Y) = D_{KL}(\mathbb{P}(X, Y), \mathbb{P}(X)\mathbb{P}(Y)) = \mathbb{E}_{X, Y} \ln \frac{\mathbb{P}(X, Y)}{\mathbb{P}(X)\mathbb{P}(Y)}$, and define the mutual information between X and Y conditioned on Z under \mathbb{P} as $I(X; Y | Z) = \mathbb{E}_{X, Y, Z} \ln \frac{\mathbb{P}(X, Y | Z)}{\mathbb{P}(X | Z)\mathbb{P}(Y | Z)}$.
- For a sequence of random variables X_1, X_2, \dots , denote by X^n the subsequence $\{X_1, X_2, \dots, X_n\}$.

We will use following two information-theoretic lower bounds.

Lemma 16. Let \mathcal{W} be a class of parameters, and $\{P_w : w \in \mathcal{W}\}$ be a class of probability distributions indexed by \mathcal{W} over some sample space \mathcal{X} . Let $d : \mathcal{W} \times \mathcal{W} \rightarrow \mathbb{R}$ be a semi-metric. Let $\mathcal{V} = \{w_1, \dots, w_M\} \subseteq \mathcal{W}$ such that $\forall i \neq j, d(w_i, w_j) \geq 2s > 0$. Let V be a random variable uniformly taking values from \mathcal{V} , and X be drawn from P_V . Then for any algorithm \mathcal{A} that given a sample X drawn from P_w outputs $\mathcal{A}(X) \in \mathcal{W}$, the following inequality holds:

$$\sup_{w \in \mathcal{W}} P_w(d(w, \mathcal{A}(X)) \geq s) \geq 1 - \frac{I(V; X) + \ln 2}{\ln M}$$

Proof. For any algorithm \mathcal{A} , define a test function $\hat{\Psi} : \mathcal{X} \rightarrow \{1, \dots, M\}$ such that

$$\hat{\Psi}(X) = \arg \min_{i \in \{1, \dots, M\}} d(\mathcal{A}(X), w_i)$$

We have

$$\sup_{w \in \mathcal{W}} P_w(d(w, \mathcal{A}(X)) \geq s) \geq \max_{w \in \mathcal{V}} P_w(d(w, \mathcal{A}(X)) \geq s) \geq \max_{i \in \{1, \dots, M\}} P_{w_i}(\hat{\Psi}(X) \neq i)$$

The desired result follows by classical Fano's Inequality:

$$\max_{i \in \{1, \dots, M\}} P_{w_i}(\hat{\Psi}(X) \neq i) \geq 1 - \frac{I(V; X) + \ln 2}{\ln M}$$

\square

Lemma 17. [Anthony and Bartlett, 2009, Lemma 5.1] Let $\gamma \in (0, 1)$, $\delta \in (0, \frac{1}{4})$, $p_0 = \frac{1-\gamma}{2}$, $p_1 = \frac{1+\gamma}{2}$. Suppose that $\alpha \sim \text{Bernoulli}(\frac{1}{2})$ is a random variable, ξ_1, \dots, ξ_m are i.i.d. (given α) $\text{Bernoulli}(p_\alpha)$ random variables. If $m \leq 2 \left\lfloor \frac{1-\gamma^2}{2\gamma^2} \ln \frac{1}{8\delta(1-2\delta)} \right\rfloor$, then for any function $f : \{0, 1\}^m \rightarrow \{0, 1\}$, $\mathbb{P}(f(\xi_1, \dots, \xi_m) \neq \alpha) > \delta$.

Next, we present two technical lemmas.

Lemma 18. [Long, 1995, Lemma 6] For any $0 < \gamma \leq \frac{1}{2}$, $d \geq 1$, there is a finite set $\mathcal{V} \in \mathbb{S}^{d-1}$ such that the following two statements hold:

1. For any distinct $w_1, w_2 \in \mathcal{V}$, $\theta(w_1, w_2) \geq \pi\gamma$;
2. $|\mathcal{V}| \geq \frac{\sqrt{d}}{2} \left(\frac{1}{2\pi\gamma} \right)^{d-1} - 1$.

Lemma 19. If $p \in [0, 1]$ and $q \in (0, 1)$, then $d_{\text{KL}}(p, q) \leq \frac{(p-q)^2}{q(1-q)}$.

Proof.

$$\begin{aligned} d_{\text{KL}}(p, q) &= p \ln \frac{p}{q} + (1-p) \ln \frac{1-p}{1-q} \\ &\leq p \left(\frac{p}{q} - 1 \right) + (1-p) \left(\frac{1-p}{1-q} - 1 \right) \\ &= \frac{(p-q)^2}{q(1-q)} \end{aligned}$$

where the inequality follows by $\ln x \leq x - 1$. □

Lemma 20. For any $0 \leq \eta < \frac{1}{2}$, $d > 4$, $0 < \epsilon \leq \frac{1}{4\pi}$, $0 < \delta < \frac{1}{2}$, for any active learning algorithm \mathcal{A} , there is a $w^* \in \mathbb{S}^{d-1}$, and a labeling oracle \mathcal{O} that satisfies η -bounded noise condition with respect to w^* , such that if with probability at least $1 - \delta$, \mathcal{A} makes at most n queries to \mathcal{O} and outputs $\hat{w} \in \mathbb{S}^{d-1}$ such that $\mathbb{P}[\text{sign}(\hat{w} \cdot x) \neq \text{sign}(w^* \cdot x)] \leq \epsilon$, then $n \geq \frac{d \ln \frac{1}{\epsilon}}{16(1-2\eta)^2}$.

Proof. We will prove this Lemma using Lemma 16.

First, we construct \mathcal{W} , \mathcal{V} , d , s , and P_θ . Let $\mathcal{W} = \mathbb{S}^{d-1}$. Let \mathcal{V} be the set in Lemma 18 with $\gamma = 2\epsilon$. For any $w_1, w_2 \in \mathcal{W}$, let $d(w_1, w_2) = \theta(w_1, w_2)$, $s = \pi\epsilon$. Fix any algorithm \mathcal{A} . For any $w \in \mathcal{W}$, any $x \in \mathcal{X}$, define $P_w[Y = 1|X = x] = \begin{cases} 1 - \eta, & w \cdot x \geq 0 \\ \eta, & w \cdot x < 0 \end{cases}$, and $P_w[Y = 0|X = x] = 1 - P_w[Y = 1|X = x]$. Define P_w^n to be the distribution of n samples $\{(X_i, Y_i)\}_{i=1}^n$ where Y_i is drawn from distribution $P_w(Y|X_i)$ and X_i is drawn by the active learning algorithm \mathcal{A} based solely on the knowledge of $\{(X_j, Y_j)\}_{j=1}^{i-1}$.

By Lemma 18, we have $M = |\mathcal{V}| \geq \frac{\sqrt{d}}{2} \left(\frac{1}{4\pi\epsilon} \right)^{d-1} - 1 \geq \frac{1}{4} \left(\frac{1}{4\pi\epsilon} \right)^{d-1}$, and $d(w_1, w_2) \geq 2\pi\epsilon = 2s$ for any distinct $w_1, w_2 \in \mathcal{V}$.

Clearly, for any $w \in \mathcal{W}$, if the optimal classifier is w , and the oracle \mathcal{O} responds according to $P_w(\cdot | X = x)$, then it satisfies η -bounded noise condition. Therefore, to prove the lemma, it suffices to show that if $n \leq \frac{d \ln \frac{1}{\epsilon}}{16(1-2\eta)^2}$, then

$$\sup_{w \in \mathcal{W}} P_w(d(w, \mathcal{A}(X^n, Y^n)) \geq s) \geq \frac{1}{2}.$$

Now, by Lemma 16,

$$\sup_{w \in \mathcal{W}} P_w^n(d(w, \mathcal{A}(X^n, Y^n)) \geq s) \geq 1 - \frac{I(V; X^n, Y^n) + \ln 2}{\ln M} \geq 1 - \frac{I(V; X^n, Y^n) + \ln 2}{(d-1) \ln \frac{1}{4\pi\epsilon} - \ln 4}.$$

It remains to show if $n = \frac{d \ln \frac{1}{\epsilon}}{16(1-2\eta)^2}$, then $I(V; X^n, Y^n) \leq \frac{1}{2} ((d-1) \ln \frac{1}{4\pi\epsilon} - \ln 4) - \ln 2$.

By the chain rule of mutual information, we have

$$I(V; X^n, Y^n) = \sum_{i=1}^n \left(I(V; X_i | X^{i-1}, Y^{i-1}) + I(V; Y_i | X^i, Y^{i-1}) \right)$$

First, we claim V and X_i are conditionally independent given $\{X^{i-1}, Y^{i-1}\}$, and thus $I(V; X_i | X^{i-1}, Y^{i-1}) = 0$. The proof for this claim is as follows. Since the selection of X_i only depends on algorithm \mathcal{A} and X^{i-1}, Y^{i-1} , for any $v_1, v_2 \in \mathcal{V}$, $\mathbb{P}(X_i | v_1, X^{i-1}, Y^{i-1}) = \mathbb{P}(X_i | v_2, X^{i-1}, Y^{i-1})$. Thus,

$$\begin{aligned} \mathbb{P}(X_i | X^{i-1}, Y^{i-1}) &= \sum_v \mathbb{P}(X_i, v | X^{i-1}, Y^{i-1}) \\ &= \sum_v \mathbb{P}(v) \mathbb{P}(X_i | v, X^{i-1}, Y^{i-1}) \\ &= \frac{1}{|\mathcal{V}|} \sum_v \mathbb{P}(X_i | v, X^{i-1}, Y^{i-1}) \\ &= \mathbb{P}(X_i | V, X^{i-1}, Y^{i-1}) \end{aligned}$$

Next, we show $I(V; Y_i | X^i, Y^{i-1}) \leq 5(1-2\eta)^2 \ln 2$. On one hand, since $Y_i \in \{-1, +1\}$, $I(V; Y_i | X^i, Y^{i-1}) \leq \ln 2$.

On the other hand,

$$\begin{aligned} &I(V; Y_i | X^i, Y^{i-1}) \\ &= \mathbb{E}_{X^i, Y^i, V} \left[\ln \frac{\mathbb{P}(V, Y_i | X^i, Y^{i-1})}{\mathbb{P}(V | X^i, Y^{i-1}) \mathbb{P}(Y_i | X^i, Y^{i-1})} \right] \\ &= \mathbb{E}_{X^i, Y^i, V} \left[\ln \frac{\mathbb{P}(Y_i | V, X^i, Y^{i-1})}{\mathbb{P}(Y_i | X^i, Y^{i-1})} \right] \\ &= \mathbb{E}_{X^i, Y^i, V} \left[\ln \frac{\mathbb{P}(Y_i | V, X^i, Y^{i-1})}{\mathbb{E}_{V'} \mathbb{P}(Y_i | V', X^i, Y^{i-1})} \right] \\ &\leq \mathbb{E}_{X^i, Y^i, V, V'} \left[\ln \frac{\mathbb{P}(Y_i | V, X^i, Y^{i-1})}{\mathbb{P}(Y_i | V', X^i, Y^{i-1})} \right] \\ &\leq \max_{x^i, y^{i-1}, v, v'} D_{\text{KL}} \left(\mathbb{P}(Y_i | x^i, y^{i-1}, v), \mathbb{P}(Y_i | x^i, y^{i-1}, v') \right) \\ &= \max_{x^i, y^{i-1}, v, v'} D_{\text{KL}} \left(\mathbb{P}(Y_i | x_i, v), \mathbb{P}(Y_i | x_i, v') \right) \\ &= \max_{x^i, v, v'} D_{\text{KL}} \left(P_v(Y_i | x_i), P_{v'}(Y_i | x_i) \right) \\ &\leq \frac{(1-2\eta)^2}{\eta(1-\eta)} \end{aligned}$$

where the first inequality follows from the convexity of KL-divergence, and the last inequality follows from Lemma 19.

Combining the two upper bounds, we get $I(V; Y_i | X^i, Y^{i-1}) \leq \min \left\{ \ln 2, \frac{(1-2\eta)^2}{\eta(1-\eta)} \right\} \leq 5(1-2\eta)^2 \ln 2$.

Therefore, $I(V; X^n, Y^n) \leq 5n(1-2\eta)^2 \ln 2$. If $n \leq \frac{d \ln \frac{1}{\epsilon}}{16(1-2\eta)^2} \leq \frac{\frac{1}{2}((d-1) \ln \frac{1}{4\pi\epsilon} - \ln 4) - \ln 2}{5(1-2\eta)^2 \ln 2}$, then $I(V; X^n, Y^n) \leq \frac{1}{2}((d-1) \ln \frac{1}{4\pi\epsilon} - \ln 4) - \ln 2$. This concludes the proof. \square

Lemma 21. For any $d > 0$, $0 \leq \eta < \frac{1}{2}$, $0 < \epsilon < \frac{1}{3}$, $0 < \delta \leq \frac{1}{4}$, for any active learning algorithm \mathcal{A} , there is a $w^* \in \mathbb{S}^{d-1}$, and a labeling oracle \mathcal{O} that satisfies η -bounded noise condition with respect to w^* ,

such that if with probability at least $1 - \delta$, \mathcal{A} makes at most n queries to \mathcal{O} and outputs $\hat{w} \in \mathbb{S}^{d-1}$ such that $\mathbb{P}[\text{sign}(\hat{w} \cdot x) \neq \text{sign}(w^* \cdot x)] \leq \epsilon$, then $n \geq \Omega\left(\frac{\eta \ln \frac{1}{\delta}}{(1-2\eta)^2}\right)$.

Proof. We prove this result by reducing the hypothesis testing problem in Lemma 17 to our problem of learning linear separators.

Fix $d, \epsilon, \delta, \eta$. Suppose \mathcal{A} is an algorithm that for any $w^* \in \mathbb{S}^{d-1}$, under η -bounded noise condition, with probability at least $1 - \delta$ outputs $\hat{w} \in \mathbb{S}^{d-1}$ such that $\mathbb{P}[\text{sign}(\hat{w} \cdot x) \neq \text{sign}(w^* \cdot x)] \leq \epsilon < \frac{1}{3}$, which implies $\theta(\hat{w}, w^*) \leq \frac{\pi}{3}$ under bounded noise condition.

Let $p_0 = \eta$, $p_1 = 1 - \eta$. Suppose that $\alpha \sim \text{Bernoulli}(\frac{1}{2})$ is an unknown random variable. We are given a sequence of i.i.d. (given α) Bernoulli(p_α) random variables ξ_1, ξ_2, \dots , and would like to test if α equals 0 or 1.

Define $e = (1, 0, 0, \dots, 0) \in \mathbb{R}^d$. Construct a labeling oracle \mathcal{O} such that for the i -th query x_i , it returns $2\xi_i - 1$ if $x_i \cdot e \geq 0$, and $1 - 2\xi_i$ otherwise. Clearly, the oracle \mathcal{O} satisfies η -bounded noise condition with respect to underlying halfspace $w^* = (2\alpha - 1)e = (2\alpha - 1, 0, 0, \dots, 0) \in \mathbb{R}^d$.

Now, we run learning algorithm \mathcal{A} with oracle \mathcal{O} . Let m be the number of queries \mathcal{A} makes, and $\mathcal{A}(\xi_1, \dots, \xi_m)$ be normal vector of the linear separator output by the learning algorithm. We define

$$f(\xi_1, \dots, \xi_m) = \begin{cases} 0 & \text{if } \mathcal{A}(\xi_1, \dots, \xi_m) \cdot e < 0 \\ 1 & \text{otherwise} \end{cases}.$$

By our assumption of \mathcal{A} and construction of \mathcal{O} , $\mathbb{P}\left(\theta(w^*, \mathcal{A}(\xi_1, \dots, \xi_m)) \leq \frac{1}{3}\pi\right) \geq 1 - \delta$, so $\mathbb{P}(f(\xi_1, \dots, \xi_m) = \alpha) \geq 1 - \delta$, implying $\mathbb{P}(f(\xi_1, \dots, \xi_m) \neq \alpha) \leq \delta$. By Lemma 17, $m \geq 2 \left\lfloor \frac{4\eta(1-\eta)}{(1-2\eta)^2} \ln \frac{1}{8\delta(1-2\delta)} \right\rfloor = \Omega\left(\frac{\eta \ln \frac{1}{\delta}}{(1-2\eta)^2}\right)$. \square